V

Vehicle Biofuels

Mark Holtzapple
Department of Chemical Engineering, Texas A&M University, College Station, TX, USA

Article Outline

Glossary

Biodiesel Methyl or ethyl ester of fatty acids.

Biomass Biological material from living or recently living organisms.

Bio-oil The liquid resulting from biomass pyrolysis.

Equivalence ratio The actual oxygen fed to the reactor divided by the stoichiometric oxygen needed for complete combustion.

Gasohol A mixture of gasoline and ethanol, typically 90% gasoline and 10% ethanol.

Lignocellulose Biomass composed of lignin, cellulose, and hemicellulose.

Oleaginous microorganisms Microorganisms that accumulate triacylglycerol (TAG) within their cells.

Saccharification Hydrolysis of polysaccharides to produce sugar.

Synthesis gas Mixture of carbon monoxide and hydrogen.

Triacylglycerol (TAG) A natural product containing three fatty acids linked to glycerol via ester bonds.

Definition of the Subject

Vehicle biofuels are solid, liquid, and gaseous fuels derived from biomass (e.g., corn, sugarcane, grasses, and wood) used for transportation (e.g., automobiles, trucks, planes, and trains). Among biofuels, liquids are the most valuable because they are easily stored, have a high energy density, and are readily metered into engines. Biofuels are infinitely renewable provided appropriate agricultural practices are followed.

Introduction

Historically, biological feedstocks have been used to power transportation. For example, early steam-powered trains were fueled by wood, Rudolph Diesel designed his engine to run on peanut oil, and Henry Ford designed the Model T to run on ethanol.

From 1880 to 1973, the price of crude oil was \$15–30/bbl (current dollars), so it was very difficult for biofuels to compete. As a consequence, during this time period, very few vehicles were powered by biofuels. However, during crisis situations, biofuels were employed. For example, during World War II, Europeans had difficulty obtaining gasoline, so millions of vehicles were adapted to run on gasified wood. The Arab Oil Embargo (1973) and Iranian Hostage Crisis (1979) caused the price of crude oil to rise to about \$90/bbl (current dollars), which precipitated interest in biofuels. In 1978, the US Energy Tax Act provided tax credits for ethanol (\$0.54/gal ethanol), which promoted the conversion of corn into biofuel. This act was designed to reduce dependence on foreign oil (36% imports in 1978) and to create a new market for corn, which had depressed prices at the time. In 2005, the US government provided a tax subsidy for biodiesel (\$1.00/gal). In 2008, the Farm Bill included a subsidy for cellulosic ethanol (\$1.01/gal).

Through these subsidies, the US government is promoting the use of biofuels for the following reasons:

Robert A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*, DOI 10.1007/978-1-4419-0851-3,

- Energy security – In 2009, the United States imported 59% of its net crude oil consumption, which makes its economy vulnerable to supply disruptions.
- Balance of trade – When crude oil sells for about $90/bbl, the United States spends about $1 billion per day on imports. If retained within the United States, this money would help develop domestic jobs.
- Rural economic development – The sale of biofuels will bring additional revenue to rural economies.
- Environment – The production of biofuels from polluting biomass feedstocks (e.g., manure, municipal solid waste) removes these hazards from the environment.
- Global warming – The combustion of biofuels does not contribute net carbon dioxide to the atmosphere; any CO_2 released from the combustion of biofuels is fixed via photosynthesis which recycles it as plant matter.

Vehicle Fuels

For transportation, liquid fuels are preferred because of their high energy density, ease of transport, and controllability; therefore, they command a premium price as described below:

- Liquid (e.g., gasoline, jet fuel, diesel) – $15 to $25/GJ.
- Gaseous (e.g., methane) – $2 to $5/GJ.
- Solid (e.g., coal) – $1 to $3/GJ.

Conventional Fuels

Gasoline is used primarily in spark-ignited piston engines (Otto cycle). It is composed of hydrocarbons containing 4–12 carbon atoms (average about 7–8). Gasoline grades are determined by the *octane number*, which characterizes the tendency of the fuel to *knock*, i.e., prematurely detonate. Gasoline-powered engines have a compression ratio (maximum: minimum gas volume) of about 7:1–12:1. During the compression of the air–fuel mixture, the temperature increases. Fragile molecules (e.g., large linear hydrocarbons) readily detonate whereas stable molecules (e.g., branched hydrocarbons, aromatics, oxygenates) resist detonation. High-compression engines are more efficient and powerful, and require high-octane fuels to prevent knocking. Because knocking pressurizes the gas at the wrong point in the engine cycle, it can rob efficiency and potentially cause damage. (Note: A fuel with an octane number of 100 has knocking characteristics identical to an idealized fuel containing 100% iso-octane and 0% *n*-heptane. Similarly, a fuel with an octane number of 0 has knocking characteristics identical to an idealized fuel containing 0% iso-octane and 100% *n*-heptane. Other fuel octane numbers correspond to other idealized mixtures.)

Jet fuel is used primarily in jet engines (Brayton cycle). It is composed of hydrocarbons containing 8–16 carbon atoms (average of about 12), which is similar to kerosene. Jet fuel formulations are adjusted to ensure the fuel does not freeze at high altitudes, where the temperature is cold. Also, the composition is adjusted to reduce smoke formation.

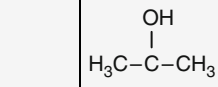
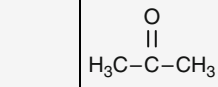
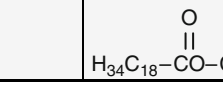
Diesel fuel is used primarily in diesel engines. It is composed of hydrocarbons containing 10–22 carbon atoms (average of about 16). Diesel engines have a high compression ratio (typically 14:1–24:1), which greatly increases the air temperature during the compression stroke. To ignite the fuel, they do not employ a spark; rather, they rely on the high temperature of the compressed air. When atomized diesel fuel is injected into the hot compressed air, it decomposes and ignites. The burn characteristics of diesel fuel are determined by the *cetane number*, which quantifies how quickly the fuel starts to auto-ignite in a diesel engine. Analogous to octane rating, the cetane rating assigns a value of 100 to a fuel that behaves like cetane (*n*-hexadecane) whereas a value of 0 is assigned to a fuel that behaves like 1-methyl napthalene.

Biofuels

Biofuels come in many forms, as described in Table 1. Table 2 provides properties of both biofuels and conventional fuels.

In the United States, ethanol is the dominant biofuel. Figure 1 shows the US ethanol capacity from 1980 to 2010. In 2010, the US ethanol capacity was 13.2 billion gallons per year and US gasoline consumption was 138.5 gal per year (Fig. 2); thus, about 8.7% of the gasoline–ethanol pool was ethanol. In 2010, approximately 40% of the US corn crop was converted to fuel ethanol. Table 3 shows global ethanol production in 2009.

Vehicle Biofuels. Table 1 Examples of biofuels

Fuel type	Generic	Examples
Primary alcohols	R–OH	$H_3C–OH$ methanol
		$H_3CH_2C–OH$ ethanol
Secondary alcohols		 isopropanol
Ethers	R–O–R'	$H_3C–O–CH_3$ dimethyl ether
		$H_3CH_2C–O–CH_2CH_3$ diethyl ether
Ketones		 Acetone
Esters		 Biodiesel

In 2009, the US biodiesel capacity was 2.0 billion gallons per year [4] and US diesel consumption was 55.7 billion gallons per year [5]; thus, about 3.5% of the diesel pool was biodiesel.

Biofuel Blends

Biofuels are often blended with other components. Common designations are described below.

E10 *E10* is a blend of 10% ethanol and 90% gasoline, and is commonly described as *gasohol*. It is widely used in the United States; nearly the entire gasoline pool contains ethanol. Modern gasoline engines are designed to be compatible with E10 without the need for modifications.

Addition of ethanol to gasoline has some benefits, such as (1) reducing gasoline imports, (2) improving the fuel octane rating, and (3) improving combustion. Regarding the latter point, the 1990 US Clean Air Act required the addition of oxygenates to *reformulated gasoline* in cities that could not attain carbon monoxide or ozone standards. In those cities, fuels were required to contain 2.7% oxygen, which could be met by adding oxygenates such as methyl tertiary butyl ether (MTBE) or ethanol. When MTBE was found to contaminate groundwater, its use was banned so the demand for ethanol increased dramatically.

Addition of ethanol to gasoline also has some detriments, such as (1) reducing fuel economy because of its lower energy content (~3%); (2) raising the fuel vapor pressure, which reduces the amount of low-cost, high-octane butanes that can be added to fuel; (3) increasing production of formaldehyde and acetaldehyde in tail-pipe emissions; and (4) absorbing water into the fuel. Regarding the latter point, ethanol is very polar compared to gasoline and attracts water, which is also polar. Common-carrier pipelines that transport hydrocarbon fuels often contain water. If ethanol-containing fuel were transported through these pipelines, it would absorb water, which adversely affects fuel quality. To avoid this problem, anhydrous ethanol is transported via rail or trucks and is "splash blended" at the fuel terminal.

E15 *E15* is a blend of 15% ethanol and 85% gasoline. As ethanol is increasingly added to the gasoline pool, the United States will soon reach a "blend wall" for E10. The US Environmental Protection Agency has authorized the use of E15 in certain vehicles (e.g., automobiles, SUVs, light-duty trucks) after model year 2001, which potentially would allow the percentage of ethanol to increase by 50%. The ruling has spurred controversy regarding potential damage to engines, emission systems, and air quality. Further, E15

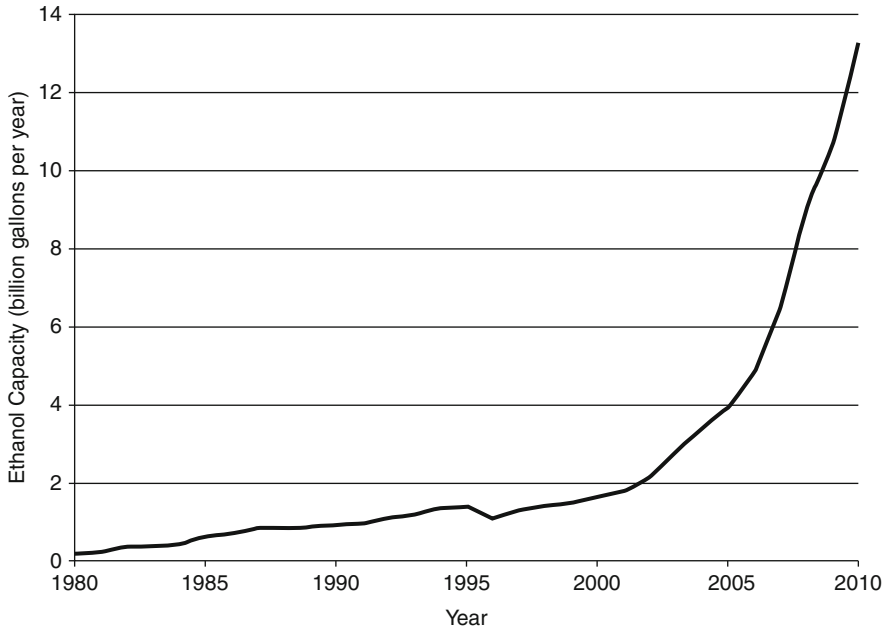
Vehicle Biofuels. Table 2 Properties of fuels[a]

Fuel	Heat of combustion[b] (MJ/kg)	Density (kg/L)	Heat of combustion[b] (MJ/L)	Octane rating[c]
Hydrocarbons				
Diesel	44.8	0.83	37.2	15–25
Jet	46.6	0.81	37.7	
Gasoline	47.3	0.74	35.0	91–99
Esters				
Biodiesel	39.9	0.88	35.1	
Alcohols				
Methanol	19.9	0.79	15.7	106
Ethanol	28.9	0.79	22.8	108
n-Propanol	30.7	0.80	24.6	
Isopropanol	30.5	0.80	24.4	
n-Butanol	33.1	0.81	26.8	96
Isobutanol	33.0	0.80	26.4	
Tertbutanol	32.6	0.78	25.4	103
n-Pentanol	34.7	0.81	28.1	
Ethers				
Dimethyl ether	28.7	0.67	19.2	
Diethyl ether	33.9	0.71	24.1	
Dipropyl ether	36.4	0.74	26.9	
Dibutyl ether	37.8	0.77	29.1	
Ketones				
Acetone	28.6	0.78	22.3	
Methyl ethyl ketone	31.5	0.80	25.2	
Diethyl ketone	35.7	0.82	29.3	

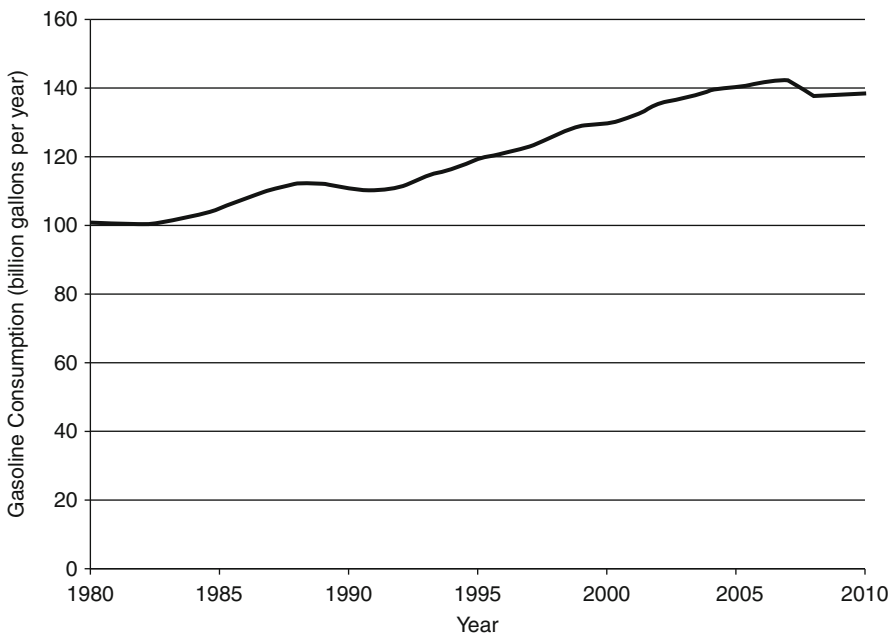
[a]Data derived from Wikipedia
[b]Higher heating value (i.e., water product is a liquid)
[c]Research Octane Number (RON)

would need to be sold in a manner to prevent it from being used in unauthorized vehicles (e.g., motorcycles, heavy-duty vehicles, off-road vehicles, and vehicles older than model year 2000).

E85 *E85* is a blend of 85% ethanol and 15% gasoline. Because ethanol has lower energy content than gasoline, the fuel mileage of E85 decreases by 25–30%. Compared to gasoline, ethanol has a higher latent heat of vaporization and lower vapor pressure; thus, it is potentially difficult to start the engine during cold weather. To overcome this problem, gasoline is a component of E85. During very cold weather, fuel retailers will increase the percentage of gasoline to 30%. Conventional spark-ignited engines are not designed to combust E85; however, flexible-fuel vehicles (FFVs) can accept any ratio of E0 to E85. In FFVs, the metals, polymers, and elastomers in the fuel system and



Vehicle Biofuels. Figure 1
US ethanol capacity [1]



Vehicle Biofuels. Figure 2
US gasoline consumption [2]

Vehicle Biofuels. Table 3 Global fuel ethanol production in 2009 [3]

Country	Annual Production (million gallons)
United States	10,600.00
Brazil	6,577.89
European Union	1,039.52
China	541.55
Thailand	435.20
Canada	290.59
India	91.67
Colombia	83.21
Australia	56.80
Other	247.27
Total	19,963.70

engine components must be compatible with the range of fuels. Further, engine sensors detect the fuel composition and automatically adjust the fuel injector and spark timing. For the automobile manufacturer, the marginal cost of producing an FFV is about $150 per vehicle.

E100 *E100* is 100% ethanol and 0% gasoline. The previously discussed fuels (E10, E15, E85) contain ethanol blended with gasoline. In these fuels, only anhydrous ethanol can be used; otherwise, water will phase out when the gasoline is added. However, E100 is a *neat fuel* (i.e., it contains no additives), so it can contain water. In the production of ethanol, it must be distilled from water. The top of the distillation column reaches the *azeotrope*, where the composition of the liquid and vapor are identical (95.63% ethanol and 4.37% water, by weight). To break the azeotrope requires additional technology (e.g., the use of molecular sieves to remove water from the solution), which adds cost. In Brazil, many vehicles operate on azeotropic ethanol; therefore, the cost of the final dehydration step is eliminated. To overcome potential cold-start problems, it is common practice to start the engine on gasoline and then switch to E100 once the engine is warmed. Direct injection of ethanol into the engine shows an effective octane rating of 130. With appropriate engine controls, it is possible to burn neat ethanol in engines with a compression ratio up to 19.5:1, which has a thermal efficiency similar to diesel engines. The higher efficiency of the high-compression engine compensates for the lower energy content of ethanol, so the fuel mileage is similar to gasoline.

Biodiesel *Biodiesel* is produced from vegetable oil or animal fats. It has a carbon number similar to petroleum-derived diesel fuel, so it can be combusted in conventional diesel engines. Common blends include B2, B5, B10, and B100, which contain 2%, 5%, 10%, and 100% biodiesel, respectively. Disadvantages of biodiesel include the following: (1) potential cold-start problems (not unlike conventional diesel), which can be overcome using additives that prevent clogging of filters; (2) deposits and clogging from low-quality or oxidized fuel; (3) slightly higher NO_x emissions; and (4) lower energy content. Regarding the latter point, because biodiesel contains oxygen, B100 has 3–5% less energy content than petroleum-derived diesel. Advantages of biodiesel include the following: (1) better lubricity (short-term studies with biodiesel show that it has less wear than conventional diesel) and (2) lower emissions. Regarding the latter point, because of its oxygen content, biodiesel burns more completely. Further, it has no sulfur or aromatic emissions.

Although biodiesel is more commonly used in diesel engines, it can be used in aircraft engines as well. Usually, it is blended with conventional jet fuel; however, there are some reports of aircraft operating with B100.

E-Diesel Because of its hydroxyl group, ethanol has hydrophilic properties and is completely soluble in water in any ratio. Because of its ethyl group, ethanol has oleophilic properties and is completely soluble in diesel fuel in any ratio. However, if a small amount of water (0.2%) is present in the fuel, two phases form: one ethanol-rich and the other diesel-rich [6]. To prevent phase separation, additives (e.g., surfactants, emulsifiers, and co-solvents) can be included in the fuel. The addition of biodiesel to petroleum-derived diesel allows significant quantities (5–15%) of ethanol to be added to the fuel [7]. Rather than adding ethanol

directly to diesel fuel, ethanol can be stored in a separate tank and introduced as a fog or vapor into the air intake of the engine.

Biomass

Biomass is biological material from living or recently living organisms. In the context of biofuels, usually biomass is used to describe plant matter such as wood, grass, agricultural residues, energy crops, and algae. The dominant components of biomass are described below.

Free Sugars

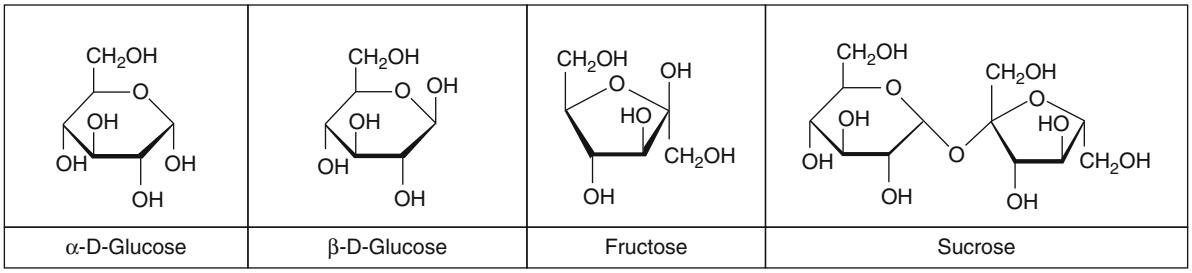
Free sugars are found in fruits and plant juices. Figure 3 shows examples including glucose, fructose, and sucrose (a disaccharide of glucose and fructose).

Starch

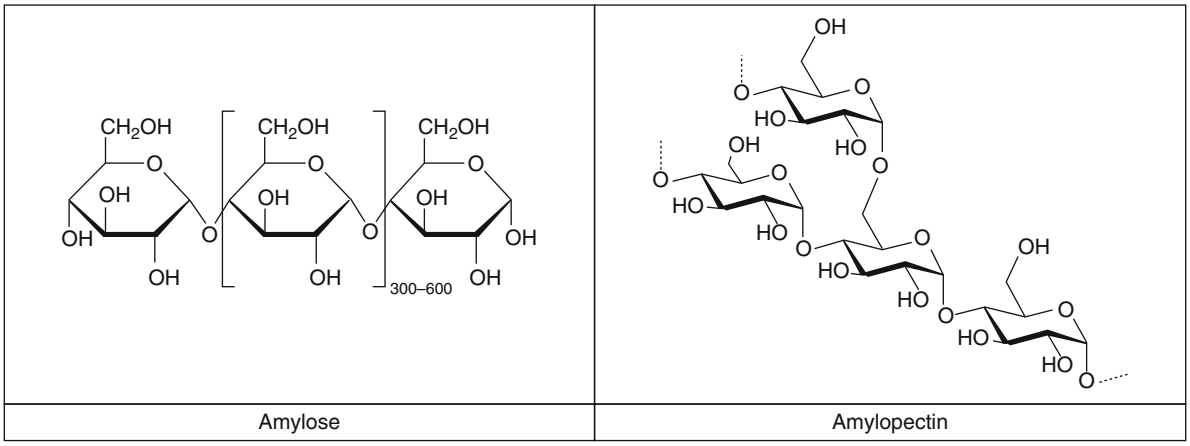
As shown in Fig. 4, starch is a polymer of glucose joined by α bonds. It occurs in two forms: amylose (unbranched) and amylopectin (branched). In plants, starch is used primarily as an energy-storage compound.

Cellulose

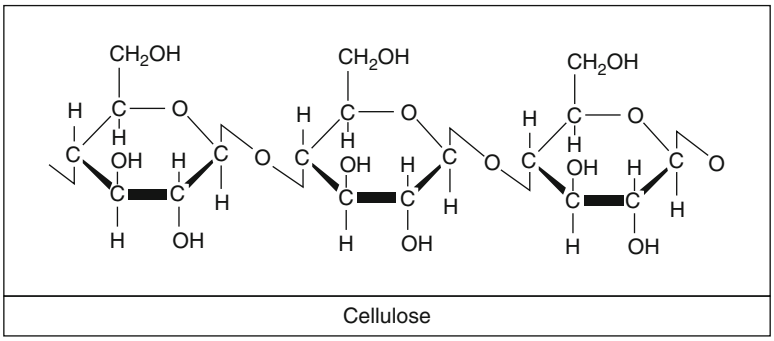
As shown in Fig. 5, cellulose is an unbranched polymer of glucose joined by β bonds. Because of extensive internal hydrogen bonds (Fig. 6), cellulose is crystalline and rigid; therefore, it is used as a structural component of roots, stems, and leaves. Because these components are so common in plants, cellulose is the most abundant biological material produced on earth.



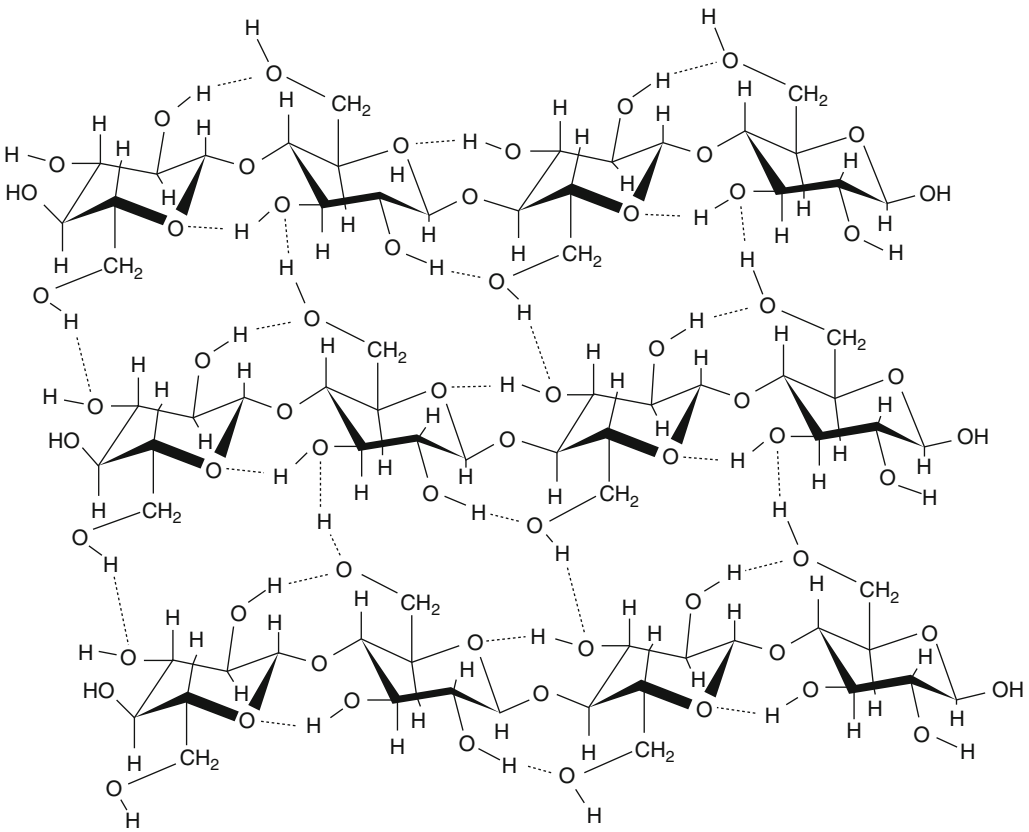
Vehicle Biofuels. Figure 3
Free sugars. (Note: For simplicity, the hydrogens bonded to carbon are not shown)



Vehicle Biofuels. Figure 4
Two forms of starch. (Note: For simplicity, the hydrogens bonded to carbon are not shown)



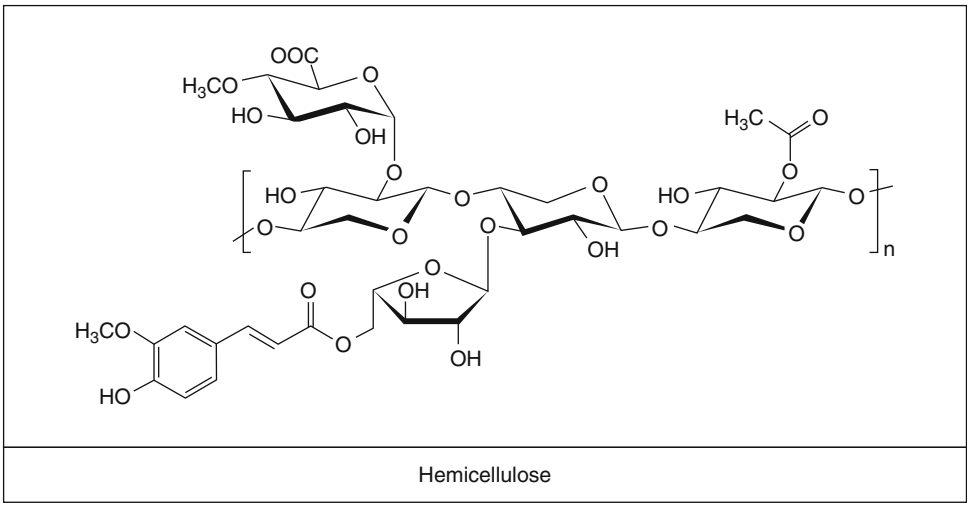
Vehicle Biofuels. Figure 5
Cellulose structure



Vehicle Biofuels. Figure 6
Cellulose has extensive hydrogen bonds, shown as *dotted lines*

Hemicellulose

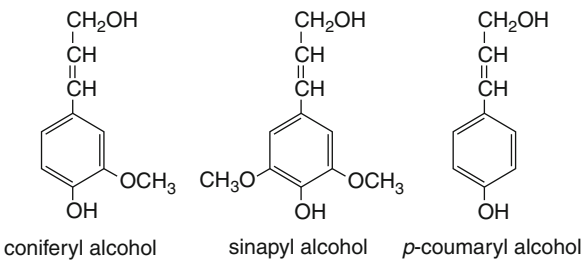
As shown in Fig. 7, hemicellulose is a polymer with a backbone of xylose (a five-carbon sugar) joined by β bonds. In addition to xylose, hemicellulose contains glucose, mannose, galactose, rhamnose, arabinose, mannuronic acid, and galacturonic acid. Hemicellulose is randomly acetylated, which helps it resist degradation by sterically hindering hemicellulose enzymes.



Vehicle Biofuels. Figure 7
Hemicellulose structure

Lignin

Lignin is a complex polymer composed of three monomers (Fig. 8), which are highly cross-linked (Fig. 9). Lignin is the "glue" that holds biomass together. By analogy to fiberglass composites, lignin functions like epoxy resin whereas cellulose functions like glass fibers. Lignin is hydrophobic and reduces evaporation of water from plant vessels. Further, lignin resists biological attack by insects and microorganisms.



Vehicle Biofuels. Figure 8
Lignin monomers

Triacylglycerol

Triacylglycerol (TAG) is the main component of vegetable oil and animal fats and is composed of fatty acids bonded to glycerol via ether linkages (Fig. 10). In Fig. 10, the value of n typically ranges from 14 (palmitic acid) to 16 (oleic and stearic acids). Although Fig. 10 shows the hydrocarbon chain is fully saturated with hydrogen, naturally occurring fatty acids often have some unsaturated bonds as well.

Proteins

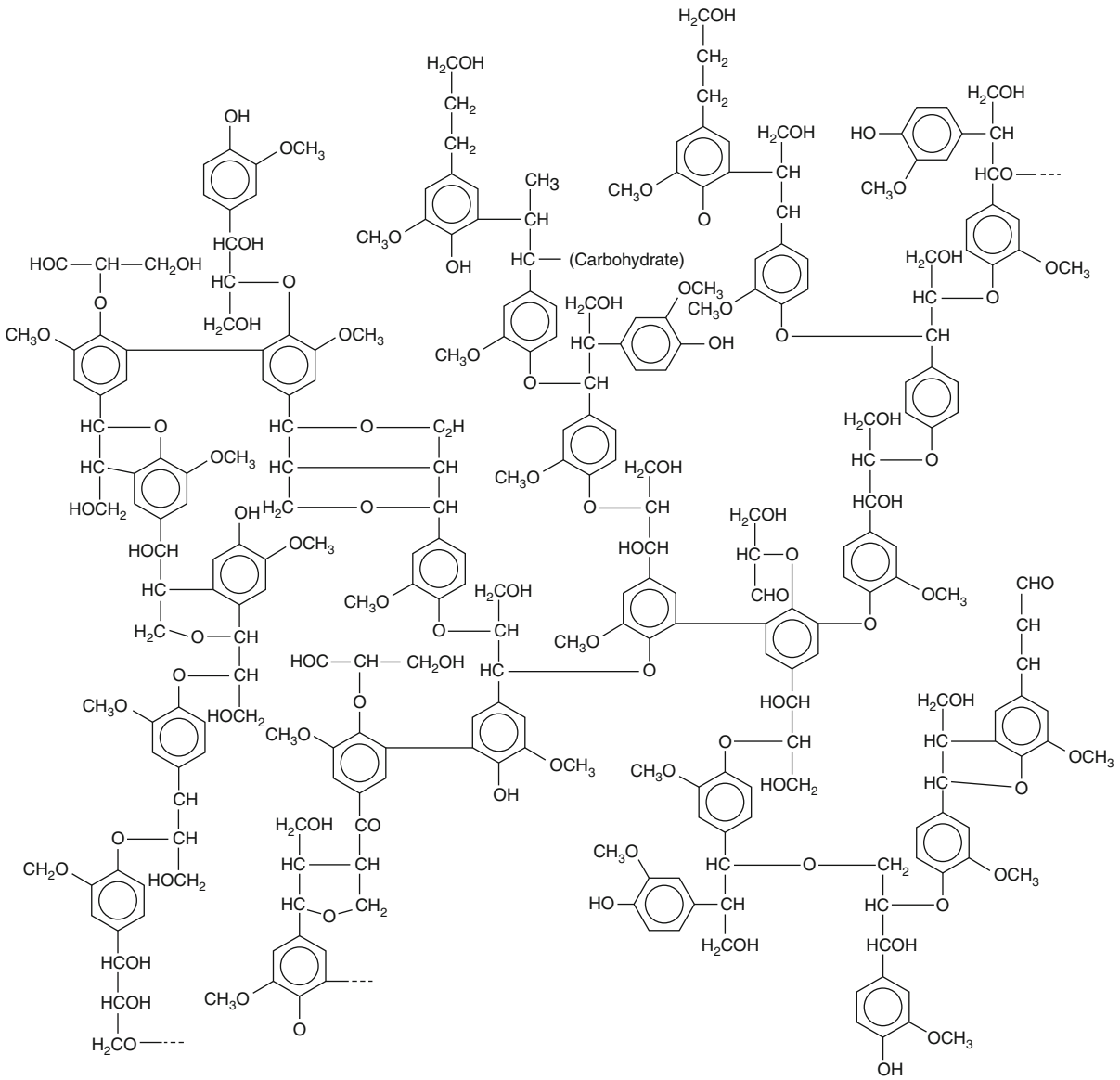
Figure 11 shows the chemical structure of an amino acid, which contains an amine group ($-NH_2$) and a carboxylic acid ($-COOH$). In naturally occurring amino acids, there are 20 standard R groups, which are coded by DNA. By eliminating water, the amine and carboxylic acid groups form a peptide bond, thus creating a protein polymer. In nature, proteins have many functions, including catalyzing reactions (enzymes) and providing mechanical structure.

Biomass Processing

Figure 12 shows an overview of processes that convert biomass into fuels, which is described in more detail below.

Raw Materials

Lignocellulose includes wood, grass, agricultural residues, and aquatic plants. Lignocellulose is the structural component of biomass and is composed of cellulose (30–50%), hemicellulose (20–40%), and



Vehicle Biofuels. Figure 9
Lignin structure

lignin (20–30%). By far, lignocellulose is the dominant form of biomass produced on earth. Table 4 catalogs the land available in the United States. Table 5 summarizes the potential lignocellulose resources in the United States. Tables 6 and 7 focus on waste biomass whereas Table 8 describes the productivity of energy crops.

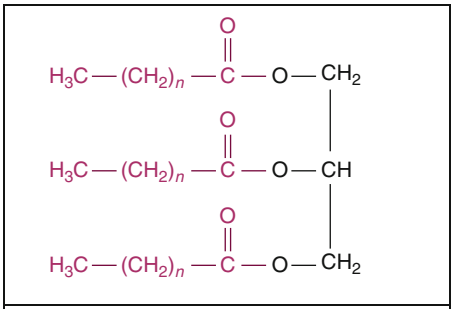
Sugar crops include sugarcane, sweet sorghum, and sugar beets. The dominant sugar is sucrose with minor amounts of glucose and fructose. Table 9 shows the productivity of sugar crops.

Starch crops include grains (e.g., corn, grain sorghum) and tubers (e.g., potatoes, cassava). Table 10 shows the productivity and abundance of US grain crops. Figure 13 shows the historical increase in US corn productivity.

Oil crops include palm, soybeans, rape, and Chinese tallow. Also, animal fats and waste frying oil can be

used as raw material. Table 11 shows the productivity of oil crops.

Algae include primarily microalgae (e.g., cyanobacteria, diatoms) but also macroalgae (e.g., kelp, seaweed). Table 12 shows the range of lipid contents and biomass productivities of microalgae grown under laboratory conditions. Table 13 shows lipid contents, biomass productivities, and lipid productivities for microalgae grown in outdoor ponds.



Vehicle Biofuels. Figure 10
Structure of triacylglycerol

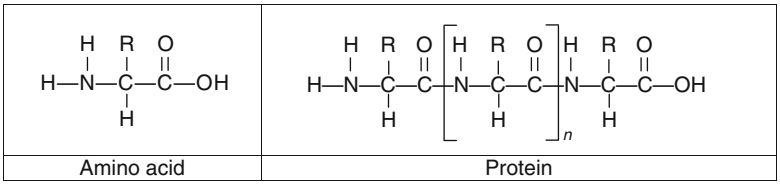
Intermediates

Figure 12 describes the following intermediates:

Bio-oil is the liquid resulting from biomass pyrolysis. Visually, it looks like crude oil; however, the composition is very different. Unlike crude oil, bio-oil contains a substantial amount of oxygen and is water soluble. A typical composition includes saccharides (2.4–3.3%), anhydrosugars (6.5–6.8%), aldehydes (10.1–14.0%), furans (0.35%), ketones (1.24–1.4%), alcohols (1.2–2.0%), carboxylic acids (8.5–11.0%), and pyrolytic lignin (16.2–20.6%). Because of the acids, the pH is low (1.5–3.8) [27].

Synthesis gas is the common name for a mixture of carbon monoxide and hydrogen, so named because it can be catalytically converted to a wide variety of products.

Acid degradation products result when biomass is heated with strong mineral acids. Components include acetaldehyde, formic acid, 2-furfural, glucose, 5-hydroxymethyl-furfural, 5-methyl-2-furfural, levulinic acid, saccharinic acid, and 2-furoic acid [28, 29].



Vehicle Biofuels. Figure 11
Structure of amino acids and proteins

Raw Material	Intermediate	Fuel
(1) Lignocellulose	(a) Bio-oil	(A) Hydrocarbon
(2) Sugar crops	(b) Synthesis gas	(B) Alcohol
(3) Starch crops	(c) Acid degradation products	(C) Ether
(4) Oil crops	(d) Sugar	(D) Ester
(5) Algae	(e) Carboxylate	(E) Ketone

Vehicle Biofuels. Figure 12
Overview of processes that convert biomass to fuels

Vehicle Biofuels. Table 4 US land categories [8]

Land categories	Percentage (%)	Land (million acre)	Land (million hectare)
Forest	33	747	302
Grassland pasture and range	26	588	238
Crop	20	453	183
Special use (e.g., public facilities)	8	181	73
Misc. (e.g., urban, swamp, desert)	13	294	119
Total	100	2,263	916

Vehicle Biofuels. Table 5 Summary of potential lignocellulose feedstocks [8]

Biomass resource	Availability (million dry ton/year)	Potential gasoline[a] (billion gal/year)
Woody		
Logging and other residues	64	4.5
Fuel treatments	60	4.2
Urban wood residues	47	3.3
Fuelwood	52	3.6
Sub-total	223	15.6
Herbaceous		
Perennial crops	377	26.4
Crop residues	446	31.2
Process residues	87	6.1
Sub-total	910	63.7
Grand total	1,133	79.3

[a]Assumed yield = 70 gal/t

Sugar intermediates include glucose, xylose, arabinose, galactose, mannose, rhamnose, and fructose, which are derived from sucrose, starch, cellulose, and hemicellulose.

Vehicle Biofuels. Table 6 Residues that are sustainably recovered under moderate crop yield increases without land use changes [8]

Biomass resource	Yield (dry ton/(acre·year))	Availability (million dry ton/year)	Potential gasoline[a] (billion gal/year)
Crop residues			
Corn	4.1	169.7	11.88
Sorghum, grain	1.7	1.3	0.09
Barley	2.2	2.8	0.20
Oats	1.9	0.7	0.05
Wheat, winter	2.3	27.4	1.92
Wheat, spring	1.4	7.4	0.52
Soybeans	1.8	0.0	0.00
Rice	5.1	10.3	0.72
Cotton lint	1.1	5.5	0.39
Other crops	1.2	20.8	1.46
Wastes			
Manure	–	43.5	3.05
Fats and grease	–	2.0	0.14
Municipal solid waste	–	29.4	2.06

[a]Assumed yield = 70 gal/t

Carboxylates are salts of carboxylic acids and include salts of volatile fatty acids (e.g., acetate, propionate, butyrate, valerate, caproate, heptanoate) and fatty acids (e.g., palmitate, stearate, oleate).

Products

Figure 12 describes the following products: hydrocarbons, alcohols, ethers, esters, and ketones. All of these products have been described previously.

Conversion Processes

Biomass is converted to fuels using either thermochemical or biological processes [30].

Vehicle Biofuels. Table 7 Maximum potential of municipal solid waste (2009) [9]

	Percentage (%)	Wet weight (mill ton/year)	Assumed moisture (%)	Dry weight (mill ton/year)	Gasoline potential (bill gal/year)
Organic					
Wood	6.5	15.8	7	14.7	1.03
Yard trimmings	13.7	33.3	70	10.0	0.70
Food scraps	14.1	34.3	80	6.9	0.48
Paper & paperboard	28.2	68.5	6	63.7	4.46
Inorganics					
Glass	4.8	11.7	0	11.7	0
Metal	8.6	20.9	0	20.9	0
Plastic	12.3	29.9	0	29.9	0
Rubber/leather/textiles	8.3	20.2	0	20.2	0
Other	3.5	8.5	0	8.5	0
Total	100.0	243.0		186.5	6.67

Vehicle Biofuels. Table 8 High-yield lignocellulosic crops

Crop	Yield (dry ton/(acre·year))	Potential gasoline[a] (gal/(acre·year))	Potential gasoline[a] (L/(ha·year))
Mixed prairie grass [10]	1.64–2.67	115–187	1,070–1,740
Switchgrass [11]	2.3–4.9	161–343	1,500–3,200
Poplar [12]	4.5–6.7	315–469	2,940–4,380
Willow [12]	4.5–6.7	315–469	2,940–4,380
Miscanthus [12]	5.3–13.4	371–938	3,460–8,760
Photoperiod-sensitive sorghum [13]	8–17	560–1,190	5,230–11,100
Conventional sugarcane [14]	17	1,190	11,100
Giant cane (*Arundo donax*) [15]	25	1,720	16,300
Energy cane [16]	30	2,100	19,600
Elephant grass (*Pennisetum purpurcum*) [17]	37–47	2,590–3,290	24,200–30,700
Water hyacinth [18]	111	7,770	72,500
Water hyacinth with enriched CO_2 [18]	146	10,220	95,400

[a]Assumed yield = 70 gal/t

Thermochemical

At elevated temperatures, thermochemical conversion processes react biomass with air or oxygen to form solid (char), liquid, and gaseous products. Except for minor amounts of nitrogen and sulfur, the elemental composition of biomass is dominated by carbon, hydrogen, and oxygen (Table 14). Free of ash, nitrogen, or sulfur, a typical formula for woody biomass is $CH_{1.4}O_{0.59}$ [31].

Vehicle Biofuels. Table 9 US sugar crops (2010)

Crop	Land harvested (acre)	Yield (wet ton/(acre·year))	Sugar content (ton sugar/t wet biomass)	Yield (ton sugar/(acre·year))	Potential ethanol yield (gal/(acre·year))
Sugarcane [8]	1,766,400	31.8	0.12	3.8	536
Energy cane [16]	Negligible	100	0.09	9	1,270
Sugar beets [8, 19]	1,155,700	27.6	0.15	4.1	578
Sorghum, sweet [20, 21]	negligible	13–20	0.06	0.78–1.2	110–169

Ton = 2,000 lb
Assumed yield = 141 gal per ton of sucrose

Vehicle Biofuels. Table 10 US grain crops (2010) [22]

Crop	Land harvested (acre)	Yield	Bushel weight (lb)	Water (%)	Yield (ton/(acre·year)	Yield (tonne/(ha·year)
Corn	81,446,000	152.8 bu/acre	56	15.5	3.62	8.11
Wheat	47,637,000	46.4 bu/acre	60	13.5	1.20	2.69
Sorghum, grain	4,808,000	71.8 bu/acre	56	13.0	1.75	3.92
Rice	3,615,000	6,725 lb/acre	NA	14.0	2.89	6.47
Oats	1,263,000	64.3 bu/acre	32	12.0	0.91	2.04

Tonne = 1,000 kg
Ton = 2,000 lb

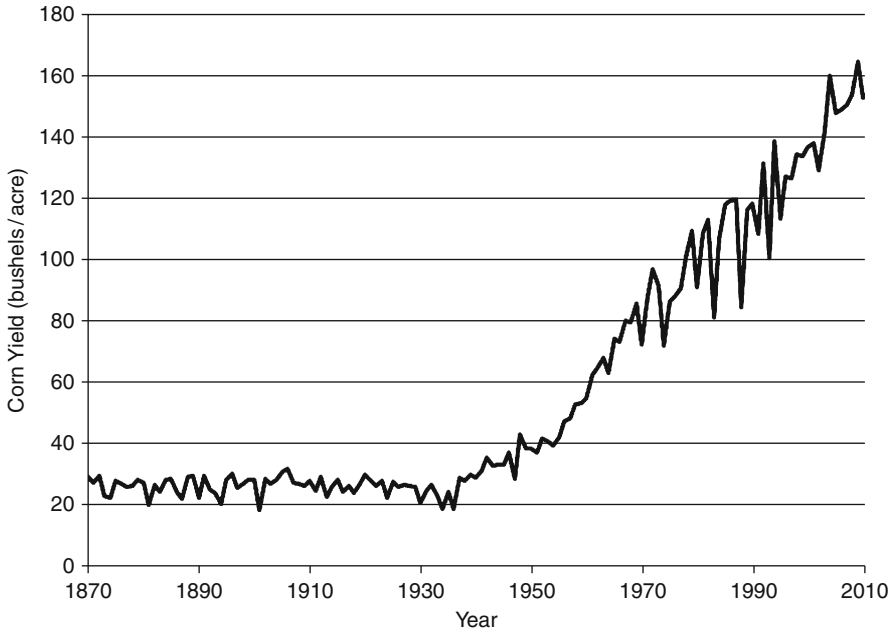
The *equivalence ratio* ϕ is defined as the actual oxygen fed to the reactor divided by the stoichiometric oxygen needed for complete combustion. Using a typical formula for woody biomass, the stoichiometric amount of oxygen is calculated as follows:

$$CH_{1.4}O_{0.59} + 1.055O_2 \rightarrow CO_2 + 0.7H_2O.$$

As a function of equivalence ratio, Fig. 14 shows the equilibrium energy content of the char and gases exiting an air-fired thermochemical reactor [31]. Under *pyrolysis* conditions ($\phi < 0.1$), char is about 40–55% of the energy content of the raw biomass with the remaining energy in the form of gases. Under *gasification* conditions ($0.2 < \phi < 0.4$), char is less than 15% of the energy content of the raw biomass; most of the energy is in the form of gases. Figure 15 shows the equilibrium composition of the gases exiting an air-fired thermochemical reactor. Other than char, the only equilibrium products are CH_4, CO, H_2, CO_2, and N_2. Real thermochemical reactors also produce complex tars and bio-oils; however, these are non-equilibrium products.

Combustion A biomass combustor can be as simple as a bonfire or a sophisticated as a fluidized bed. The objective of the combustion process is to produce heat (Table 14), which typically is used to generate steam. To ensure that there is negligible carbon monoxide in the combustion gas, a biomass combustor is typically operated at an equivalence ratio of 1.05–1.10. In an air-fired combustor, the combustion temperature is about 2,050°C and in an oxygen-fired combustor, combustion temperature is about 2,800°C [31].

As a solid fuel, biomass competes with coal. Because of its lower oxygen content, coal has higher energy



Vehicle Biofuels. Figure 13
Historical corn productivity [22]

content than biomass. Heating biomass in an oxygen-free environment drives off volatiles to produce charcoal, which has lower oxygen content than biomass and hence its heating value is similar to coal. Compared to coal, biomass has fewer impurities (N, S), so less treatment of stack gases is required, which lowers the capital cost of the combustor.

Gasification Commonly, gasifiers are operated at an equivalence ratio of about 0.25 (1.6 g air/g biomass). Gasification typically occurs at temperatures of 700–1,100°C with air and about 100°C higher with oxygen [31]. Table 15 shows the energy content of the gases exiting a gasifier fired with air or oxygen. In Europe during World War II, the low-energy gas from air-fired gasifiers was fed to millions of vehicles.

The mechanical design of gasifiers follows [31]:

Updraft gasifiers employ a bed of biomass in which fresh biomass is introduced at the top. Air (or oxygen) is introduced at the bottom of the bed and is blown upward. The bottom zone is richest in oxygen, which is where combustion reactions occur thereby releasing heat and increasing the temperature. As the hot gases flow upward through the upper zone of fresh biomass, gases and liquids are generated. Ideally, to prevent liquid condensation, the hot products are burned directly.

Downdraft gasifiers also employ a bed of biomass in which fresh biomass is introduced at the top. Air (or oxygen) is introduced at an intermediate zone in the bed. The gas flows downward with the biomass and passes through the hot combustion zone. Because the gas exits at very high temperatures, less tar and bio-oil is produced. The hot product gas can be used to preheat the incoming biomass feed.

Fluidized-bed gasifiers mix high-velocity air and steam with the biomass to create an ebullating bed. Unlike fixed-bed gasifiers, fluidized-bed gasifiers can employ a wider range of particle sizes and have a higher throughput because of the uniformly high temperature. To maintain high velocities, the hot product gas is often recycled into the fluidized-bed gasifier. These gasifiers often employ a solid heat transfer agent, such as sand or catalyst.

Vehicle Biofuels. Table 11 Productivity of oil-producing crops

Crop	Latin Name	Productivity (gal/(acre·year))	Productivity (L/(ha·year)
Corn [23]	*Zea mays*	18	168
Soybean [23]	*Glycine max*	46	429
Peanut [23]	*Ariachis hypogaea*	109	1,020
Rape seed [23]	*Brassica napus*	122	1,140
Castor bean [23]	*Ricinus communis*	145	1,350
Jojoba [23]	*Simmondsia chinensis*	186	1,740
Jatropha [23]	*Jatropha curcas*	194	1,810
Coconut [23]	*Cocos nucifera*	276	2,580
Macauba palm [23]	*Acrocomia aculeata*	461	4,300
Oil palm [23]	*Elaeis guineensis*	610	5,700
Chinese tallow [24]	*Triadica sebifera*	645	6,020

Vehicle Biofuels. Table 12 Productivity of microalgae [25]

Algae species	Lipid content (%)	Biomass productivity (tonne/(ha·year))
Ankistrodesmus sp.	24–31	42–63
Botryoccus braunii	25–75	11
Chlorella emersonii	25–63	3.3–3.5
Chlorella vulgaris	5–58	2.1–3.5
Chlorella sp.	10–48	5.9–91
Chlorella pyrenoidosa	2	264–474
Chlorella	18–57	13–51
Dunaliella salina	6–25	5.8–131
Dunaliella primolecta	23	51
Haematococcus pluvialis	25	37–133
Monallanthus salina	20–22	43
Nannochloropsis sp.	12–53	7–19
Nitzchia sp.	16–47	32–79
Oocystis pusilla	11	148–167
Phaeodactylum tricornutum	18–57	8.8–77
Porphyridium cruentum	9–61	91
Scenedesmus sp.	20–21	8.9–49
Spirulina platensis	4–17	1.5–186
Spirulina maxima	4–9	91
Tetraselmis suecica	8–23	69

The primary product from gasifiers is a mixture of carbon monoxide and hydrogen, often called *synthesis gas*. The hydrogen content of the synthesis gas can be enriched using the *shift reaction*

$$CO + H_2O \rightarrow CO_2 + H_2.$$

Commonly, the process employs two stages: high-temperature shift (350°C, iron oxide catalyst) and low-temperature shift (190–210°C, copper catalyst).

Synthesis gas can be converted to a variety of products, as described below.

Methanol ($1 \rightarrow b \rightarrow B$) (*Note:* Codes are described in Fig. 12) Using ZnO catalyst operating at 240–400°C and 50–300 atm, methanol can be produced from synthesis gas according to the following reaction [31]:

$$CO + 2H_2 \rightarrow CH_3OH.$$

Mixed Alcohols ($1 \rightarrow b \rightarrow B$) Using Mo_2C catalyst operating at 300°C and 80 atm, mixed alcohols (e.g., C2 to C7) can be produced from synthesis gas [32]. A representative reaction is the formation of ethanol [32]

$$CO + 4H_2 \rightarrow C_2H_5OH + H_2O.$$

Hydrocarbons ($1 \rightarrow b \rightarrow A$) Fischer–Tropsch catalysts (e.g., cobalt, iron) typically operate at 150–300°C and produce hydrocarbons according to the following reaction:

$$(2n+1)H_2 + nCO \rightarrow C_nH_{(2n+2)} + nH_2O.$$

Vehicle Biofuels. Table 13 Productivity of microalgae grown in outdoor ponds [26]

Algae species	Lipid content (%)	Biomass productivity (tonne/(ha·year)	Lipid productivity (L/(ha·year))	Lipid productivity (gal/(acre·year))
Amphora	40	142	64,500	6,910
Chaetoceros muelleri	26	95	28,100	3,010
Cyclotella cryptica	24	99	27,000	2,890
Isochrysis galbana	22	102	25,500	2,730
Nannochloropsis	21	55	13,100	1,410
Nannochloropsis salina	16	91	16,500	1,770
Tetraselmis suecica	22	69	17,200	1,850

The product distribution tends to be very wide and includes heavy waxes. Although waxes are valuable products, the market is small, so they often must be refined into lighter fuels (e.g., gasoline, jet fuel, diesel). During World War II, Fischer–Tropsch chemistry was used by the Germans to produce liquid fuels from coal. Currently, South African SASOL uses Fischer–Tropsch chemistry to produce hydrocarbons from coal. A number of commercial plants (Malaysia, Qatar) use Fischer–Tropsch chemistry to produce hydrocarbons from natural gas.

Hydrocarbons (1 → b → B → A) Through an alternative route, methanol (or higher alcohols) can be converted to hydrocarbons using zeolite catalysts, such as ZSM-5 [31].

$$n\mathrm{CH_3OH} \rightarrow \mathrm{C}_n\mathrm{H}_{2n} + n\mathrm{H_2O}$$

Ethers (1 → b → B → C) Methanol can also be converted to dimethyl ether (DME) using a zeolite catalyst, such as ZSM-5 [33].

$$2\mathrm{CH_3OH} \rightarrow \mathrm{H_3COCH_3} + \mathrm{H_2O}$$

DME can be used in diesel engines as a clean-burning fuel [34].

Pyrolysis (1 → a → A) As stated previously, liquids are not thermodynamically stable products from thermochemical reactors; however, they are produced as reaction intermediates. *Fast pyrolysis* involves rapidly heating biomass (less than 1–5 s) to 450–600°C and then rapidly cooling the products to "freeze" the reaction [35]. On a weight basis, a typical product spectrum consists of the following:

- Char = 10–25%
- Gas = 8–30%
- Bio-oil = 45–65%
- Water = 8–15%

Although the bio-oil physically looks like crude petroleum, its properties are very different [36]. It has high water content (15–30%) that is difficult to remove via distillation. Further, it has high oxygen content (35–40%) and low pH (~2.5). Crude bio-oil can be used directly in boilers and engines (e.g., diesel, gas turbines); however, there are issues with fuel stability and corrosiveness. Using processes similar to those used to refine crude oil (e.g., hydrotreating and catalytic vapor cracking), the bio-oil can be upgraded to products similar to those obtained from petroleum [36]. At the moment, upgrading bio-oil is expensive and energy yields are low.

Acid Degradation (1 → c → C) To produce fuels using acid degradation, lignocellulose is thermochemically treated in a multi-step process [37]:

- *Step 1* – The biomass is hydrolyzed at 210–230°C for 13–25 s using 1–5% mineral acid. This produces primarily hydroxymethylfurfural, which is removed continuously.
- *Step 2* – The hydroxymethylfurfural is hydrolyzed further at 195–215°C for 15–30 min to produce primarily levulinic acid, which is continuously removed.

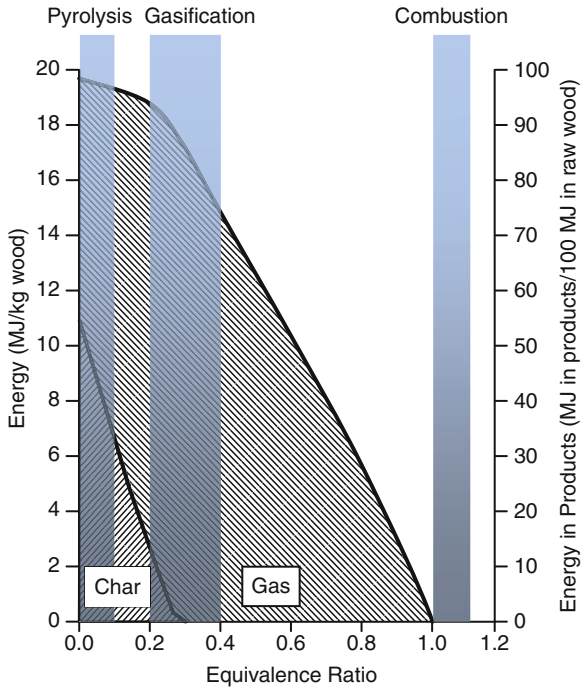
Vehicle Biofuels. Table 14 Elemental composition and energy content of solid fuels [31]

Feedstock	Elemental composition (wt%)						Higher heating value[a]	
	C	H	N	S	O	ash	(Btu/lb)	(MJ/kg)
Coal								
Pittsburg seam coal	75.5	5.0	1.2	3.1	4.9	10.3	13,650	31.754
West Kentucky No. 11 coal	74.4	5.1	1.5	3.8	7.9	7.3	13,460	31.312
Utah coal	77.9	6.0	1.5	0.6	9.9	4.1	14,170	32.963
Wyoming Elkol coal	71.5	5.3	1.2	0.9	16.9	4.2	12,710	29.567
Lignite	64.0	4.2	0.9	1.3	19.2	10.4	10,712	24.919
Charcoal	80.3	3.1	0.2	0.0	11.3	3.4	13,370	31.102
Biomass								
Douglas fir	52.3	6.3	0.1	0.0	40.5	0.8	9,050	21.053
Douglas fir bark	56.2	5.9	0.0	0.0	36.7	1.2	9,500	22.010
Pine bark	52.3	5.8	0.2	0.0	38.8	2.9	8,780	20.425
Western hemlock	50.4	5.8	0.1	0.1	41.4	2.2	8,620	20.052
Redwood	53.5	5.9	0.1	0.0	40.3	0.2	9,040	21.030
Beech	51.6	6.3	0.0	0.0	41.5	0.6	8,760	20.378
Hickory	49.7	6.5	0.0	0.0	43.1	0.7	8,670	20.169
Maple	50.6	6.0	0.3	0.0	41.7	1.4	8,580	19.959
Poplar	51.6	6.3	0.0	0.0	41.5	0.6	8,920	20.750
Rice hulls	38.5	5.7	0.5	0.0	39.8	15.5	6,610	15.377
Rice straw	39.2	5.1	0.6	0.1	35.8	19.2	6,540	15.214
Sawdust pellets	47.2	6.5	0.0	0.0	45.4	1.0	8,814	20.504
Paper	43.4	5.8	0.3	0.2	44.3	6.0	7,572	17.615
Redwood wood waste	53.4	6.0	0.1	0.1	39.9	0.6	9,163	21.316
Alabama oak waste wood	49.5	5.7	0.2	0.0	41.3	3.3	8,266	19.229
Animal waste	42.7	5.5	2.4	0.3	31.3	17.8	7,380	17.168
Municipal solid waste	47.6	6.0	1.2	0.3	32.9	12.0	8,546	19.880

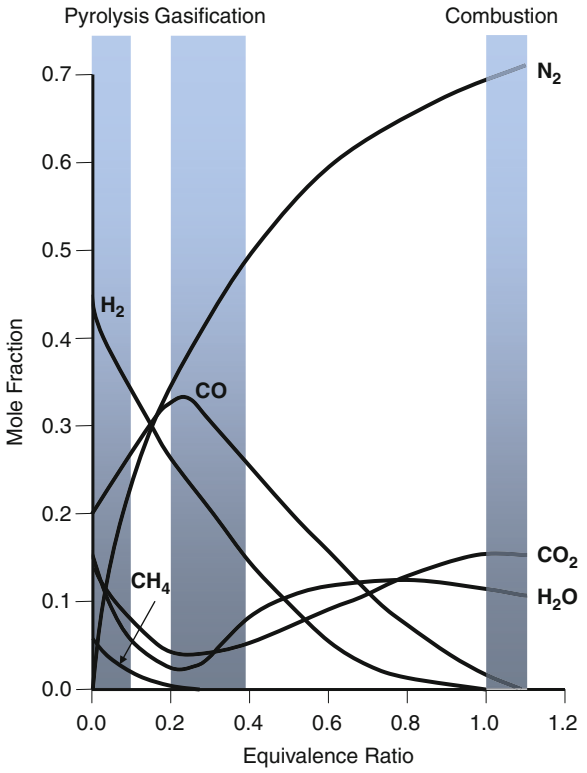
[a]Water product is assumed to be liquid

- *Step 3* – Through a series of dehydration and reduction reactions, levulinic acid is converted to methyltetrahydrofuran, an oxygenate that can be added to conventional gasoline.

Biodiesel (4 → e → D or 5 → e → D) A component of oil seeds and algae, triacylglycerol (TAG) can be extracted using a solvent (e.g., hexane). TAG can be used directly in diesel engines; however, its high viscosity causes operational problems in the fuel injectors. To overcome this problem, the diesel engine can be modified to include fuel-line heaters that lower the TAG viscosity. Rather than modify the engine, the fuel itself can be modified to lower its viscosity. Typically, this is accomplished by converting TAG to esters ("biodiesel") by a base-catalyzed reaction with an



Vehicle Biofuels. Figure 14
Equilibrium energy content of gas and char exiting a thermochemical reactor using air as the oxidant. The equivalence ratio is the actual air employed compared to the stoichiometric amount needed for complete combustion [31]



Vehicle Biofuels. Figure 15
Equilibrium gas composition exiting a thermochemical reactor using air as the oxidant. The equivalence ratio is the actual air employed compared to the stoichiometric amount needed for complete combustion [31]

alcohol [38, 39]. Because of its low cost, methanol is the preferred alcohol; however, other alcohols (e.g., ethanol) can be used as well. A by-product of biodiesel production is glycerol.

Biological Processes

Biological processing of biomass involves the use of fermentation to convert biomass to fuels.

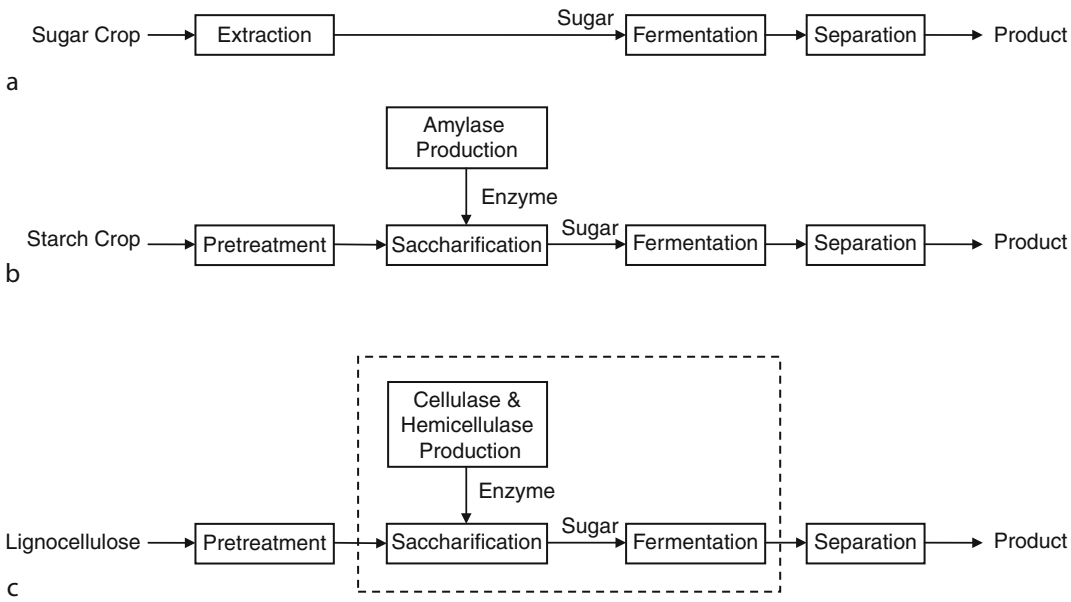
Sugar Platform The sugar platform employs processes that convert sugar to fuel (typically alcohol) via fermentation (Fig. 16).

Sugar Crop (2 → d → B) In Brazil, the dominant sugar crop is sugarcane (*Saccharum officinarum*). Traditionally, prior to harvest, the sugarcane field is set afire to burn off leaves, which contain negligible amounts of sugar. (*Note:* Because of the pollution that results from burning sugarcane fields, many governments are banning this practice.) After the fire, the stalks are brought to the sugar mill for processing. On a wet basis, the harvested plant contains 68–72%

Vehicle Biofuels. Table 15 Energy content of gaseous fuels [31]

Name	Source	Energy content (Btu/scf)[a]
Low-energy gas	Air gasifier	150–200
Medium-energy gas	Oxygen gasifier Pyrolyzer	300–500
Biogas	Anaerobic digestor	600–700
High-energy gas	Natural gas	1,000

[a]*scf* standard cubic foot (ft^3) where $T = 60°F$ and $P = 1$ atm



Vehicle Biofuels. Figure 16
Sugar platform. (**a**) sugar crop; (**b**) starch crop; (**c**) lignocellulose

moisture and 12–17% total sugars that are 90% sucrose and 10% glucose or fructose [40].

To extract the sugar, the sugarcane stalks are chopped using a hammer mill and washed countercurrently with water in a multi-stage operation. The most common extraction method uses roller mills that intensely squeeze the fiber to recover the maximum juice per stage. Typically three to four stages are employed. Another common extraction method employs diffusers in which the fiber is placed on a conveyor belt and is repeatedly washed in a countercurrent manner with water that has ever-lower sugar concentrations. Typically, 10–18 washing stages are employed [40]. The extraction process usually recovers >95% of the sugar in fiber. The washed fiber is called *bagasse*, and is used to fuel boilers that power the sugar mill.

Commercial sugarcane varieties are bred to have high sugar concentrations, which reduces the amount of fiber that must be processed by the sugar mill. In contrast, energy cane varieties are bred to have high per-acre yields of both fiber and sugar, at the expense of lower sugar concentrations. To economically recover sugar from energy cane, a novel screw-press conveyor extraction system can be employed that uses a "gentle squeeze" between extraction stages [41]. Typically eight extraction stages would be employed.

In Brazil, both continuous and batch fermentations are employed [42]. Continuous fermentations are more productive; however, they are more prone to contamination, thus batch fermentations are preferred. In both processes, centrifuges recover yeast (*Saccharomyces cerevisiae*) from the fermentation broth and recycle it to subsequent fermentations. Yeast grow at acid pH whereas bacteria prefer neutral pH; therefore, to reduce bacterial contamination, the recycled yeast is contacted with dilute sulfuric acid. Typical fermentations employ the following conditions: cell density 8–17% v/v, temperature 33–35°C, ethanol concentration 8–11% v/v, ethanol yield 90–92% of theoretical, and fermentation time 6–10 h. Contaminants are unable to take over the fermentation because of the short residence time, high ethanol concentrations, and addition of antibiotics.

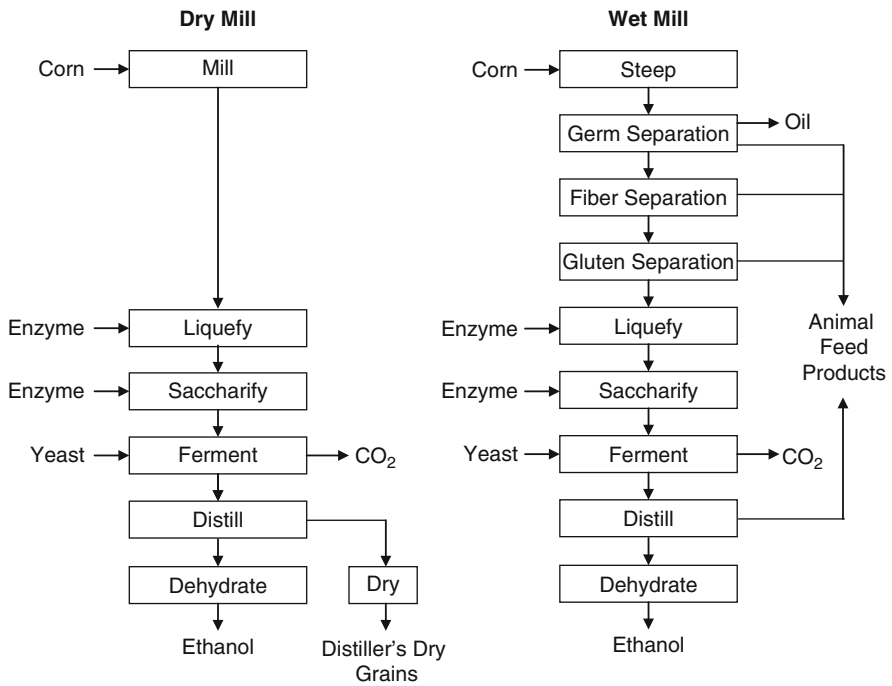
The ethanol product is distilled to the azeotrope (95.5% v/v) and then can be dehydrated using molecular sieves [42]. In Brazil, azeotropic ethanol is sold as a neat fuel; however, it cannot be blended into gasoline or diesel fuel because the water will form a separate phase.

Starch Crop ($3 \rightarrow d \rightarrow B$) The most common starch crop is corn, which has a typical composition shown in Table 16. Corn is processed by wet mills and dry mills (Fig. 17). Wet mills fractionate corn into oil, fiber, gluten (i.e., protein), and starch; the later which may be fermented into ethanol. Purpose-built ethanol plants use dry milling, which is less capital intensive and produces the majority of corn ethanol in the United States.

Vehicle Biofuels. Table 16 Typical corn composition [43]

Component	Wet basis (g/100 g wet corn)	Dry basis (g/100 g dry corn)
Starch	59.5	70.0
Fiber	7.0	8.2
Ash and other	6.7	7.9
Protein	8.4	9.9
Oil	3.4	4.0
Water	15.0	0.0
Total	100.0	100.0

In dry mills [43], the corn is hammer-milled and then is mixed with water, lime, and ammonia at 88–110°C for 75 min. During this cooking process, the starch hydrates and swells. To reduce the viscosity, thermo-stable α-amylase is added which begins the *saccharification* (i.e., sugar-production) process. After cooking, the temperature is lowered to 61°C and glucoamylase is added, which completes the saccharification of starch to glucose in about 5 h. After saccharification, the temperature is lowered to 32°C to begin the fermentation using yeast (*Saccharomyces cerevisiae*). Typically, batch fermentations are used with a residence time of about 68 h. The final ethanol concentration in the beer is about 10.8% ethanol (w/w). To recover the ethanol, the beer is distilled. The bottoms of the first distillation column contain solids (e.g., yeast, gluten, fiber, and germ), which are recovered by centrifugation and evaporation and sold as distiller's dried grains with solubles (DDGS). The DDGS has high protein content (about 28%) and is



Vehicle Biofuels. Figure 17
Corn-to-ethanol processes

sold as animal feed. The tops from the first distillation column contain ethanol and water, which is further distilled and finally dehydrated using molecular sieves.

In wet mills [44], first the corn is *steeped*, i.e., soaked in water containing 0.06–0.2% sulfur dioxide at about 51°C for about 36 h. The steeping softens the kernels and breaks the protein matrix allowing the corn to be separated into its various components (germ, fiber, gluten, starch). From the germ, corn oil is extracted leaving germ meal. The germ meal, corn fiber, and gluten are formulated into various animal feed products (e.g., gluten meal, gluten feed). The corn starch has numerous uses (e.g., production of high-fructose syrup) and can be used to produce ethanol in a process similar to that used in dry mills.

Although ethanol is the dominant alcohol produced from starch, butanol is also being produced [45].

Lignocellulose (1 → d → B) In lignocellulose, the dominant polysaccharide is cellulose, which is composed of β-linked glucose monomers. Because of its similarity to starch (α-linked glucose), many lignocellulose-conversion processes are modeled after starch-conversion processes (Fig. 16). In starch-conversion processes, the pretreatment step involves soaking in water, which swells the starch rendering it more susceptible to enzymatic hydrolysis. In lignocellulose-conversion processes, a more aggressive pretreatment is required.

The enzymatic digestibility of lignocellulose is affected by the following factors: lignin content, acetyl content of hemicellulose, cellulose crystallinity, degree of polymerization, pore volume, and accessible surface area. To increase its enzymatic digestibility, lignocellulose may be treated with steam, alkali (e.g., ammonia, lime, sodium hydroxide), dilute acids (e.g., sulfuric), liquid hot water, solvents (e.g., ethanol, acetone, butanol, ionic liquids), oxidants (e.g., hydrogen peroxide, oxygen, ozone), and machines (e.g., ball mill, two-roll mill) [46].

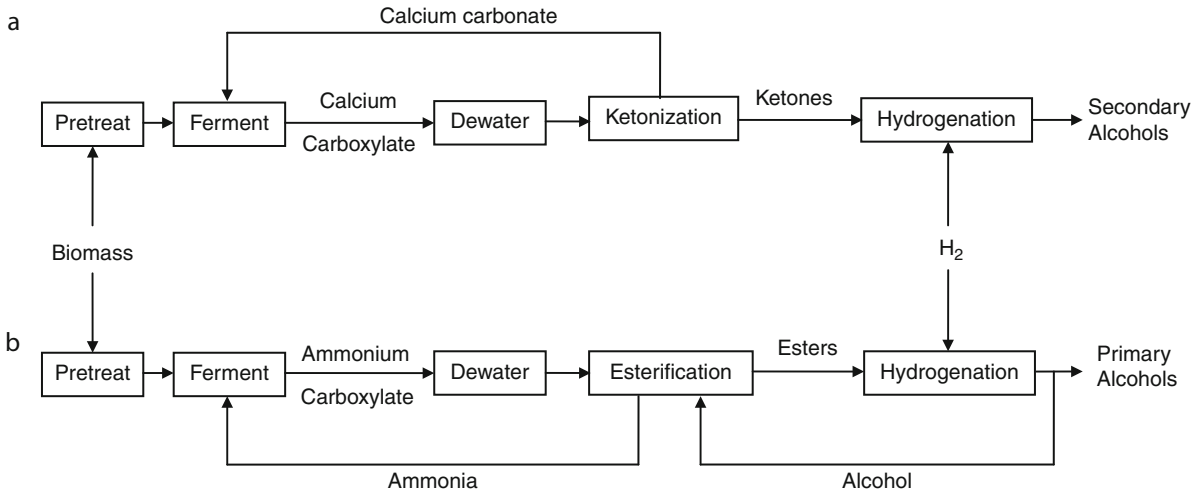
After pretreatment, the lignocellulose is saccharified using cellulase and hemicellulase enzymes, which produce primarily glucose and xylose [47]. Traditional yeast (*Saccharomyces cerevisiae*) can ferment only the glucose, so xylose-fermenting microorganisms (e.g., *Pichia stipitis*) are required to utilize all sugars. Using genetic engineering, microorganisms have been developed that ferment both glucose and xylose to ethanol [48]. Economies can be realized via *consolidated bioprocessing* in which enzyme production, saccharification, and fermentation occur in the same vessel using a single microorganism, as shown in the dotted box of Fig. 16 [49]. Although genetic engineering is being used to develop consolidated bioprocessing microorganisms, naturally occurring microorganisms have been identified as well [50].

The conversion of lignocellulose to ethanol has not been commercialized, so typical industrial fermentation conditions have not yet been established. Compared to starch, lignocellulose is much less reactive; therefore, lignocellulose fermentations are likely to take much longer and produce lower ethanol concentrations.

Lignocellulosic ethanol will be recovered via distillation, much like ethanol from conventional starch or sugar processes.

The main impediments to lignocellulosic ethanol are the high cost of pretreatment and enzymes. Some processes overcome this barrier by using dilute acids to saccharify lignocellulose [51]. In the early and mid twentieth century, this process was used commercially in Germany and the former Soviet Union, and was piloted in the United States during World War II. The main obstacle to widespread commercialization is sugar degradation, which reduces yields and inhibits the fermentation. Another process option is the use of concentrated acid to saccharify lignocellulose [52]. Compared to dilute-acid hydrolysis, sugar degradation is much less; however, acid recovery is a challenge. One promising approach is industrial-scale chromatography.

Carboxylate Platform (1 → e → B) The carboxylate platform is a type of consolidated bioprocessing in which lignocellulose is biologically converted to carboxylate salts (e.g., calcium acetate) that are subsequently chemically converted to fuels and chemicals [53, 54]. Generally, a mixed culture of microorganisms is used, which has a great operational advantage: sterile operating conditions are not required. Nonetheless, because they have greater control over the products and yields, some processes use a monoculture and do require sterile operating conditions [55].



Vehicle Biofuels. Figure 18
Carboxylate platform. (**a**) secondary alcohol route; (**b**) primary alcohol route

Figure 18 shows a schematic of the carboxylate platform, which has routes to secondary and primary alcohols.

In the secondary alcohol route [56], the biomass is pretreated to enhance its digestibility. In principle, any pretreatment can be used; however, alkaline pretreatments (e.g., lime) are favored because the cations help neutralize the carboxylic acids produced in the fermentation. The biomass is then fed to a mixed culture of microorganisms that ferments biomass components (e.g., cellulose, hemicellulose, starch, free sugars, pectin, protein, fats) into carboxylic acids (e.g., acetic, propionic, butyric acids) and hydrogen. To maintain a near-neutral pH, a buffer is added (e.g., calcium carbonate). The mixed culture of microorganisms can originate from a variety of sources, such as cattle rumen or marine swamps. An inhibitor (e.g., iodoform) is added to suppress the generation of methane, an unwanted by-product. Via vapor-compression evaporation, the carboxylate salts in the fermentation broth are dewatered. The concentrated carboxylate salts (e.g., calcium acetate) are thermally converted to ketones (e.g., acetone). The ketones are hydrogenated using a hydrogenation catalyst (e.g., Raney nickel) to produce secondary alcohols (e.g., isopropanol). Potential sources of hydrogen include the fermentor gas, gasified undigested residues, reformed methane, and water electrolysis.

In the primary alcohol route [57], the early steps (pretreatment, fermentation, and dewatering) are nearly identical to the secondary alcohol route. After dewatering, the concentrated carboxylate salts (e.g., ammonium acetate) react with a recycled high-molecular-weight alcohol (e.g., hexanol) to form esters (e.g., hexyl acetate), which is subsequently hydrogenated to form two alcohols (e.g., ethanol and hexanol). The high-molecular-weight alcohol is recycled and the low-molecular-weight alcohol is harvested as product.

Syngas Platform (1 → b → B) As described previously, syngas ($CO + H_2$) is produced by gasifying biomass. The syngas can be fermented to acetate, butyrate, ethanol, and butanol [58–60].

Biomass to Hydrocarbons Because of their high energy density, hydrocarbons are the preferred source of transportation fuels. Various routes to hydrocarbons are described below.

Biogas (1 → e → A) As described previously, the carboxylate platform uses a mixed culture of microorganisms to produce carboxylate salts and hydrogen. If no methanogen inhibitor is included, these products are converted to biogas ($CH_4 + CO_2$), which simply bubbles out of the fermentation broth [61].

Because of its simplicity, this process is used all over the world, including developing countries, to convert waste biomass into fuel. Normally, this fuel is used for stationary applications (e.g., cooking, electricity production); however, it could be purified by removing carbon dioxide using acid-gas absorbents (e.g., triethanol amine). The resulting pure methane could be cryogenically liquefied and used for transportation purposes.

Hydrocarbons from Alcohols (B → A) Fermentation-derived alcohols can be converted to hydrocarbons using zeolite catalysts, such as H-ZSM-5 [62].

$$n\mathrm{C_2H_5OH} \rightarrow \mathrm{C_{2n}H_{4n}} + n\mathrm{H_2O}$$

The reaction proceeds by dehydrating the alcohol to form an alkene (olefin), which then oligomerizes to form higher molecular weight hydrocarbons (e.g., olefins, aromatics). If desired, the final product can be hydrogenated to saturate the bonds. Compared to low-molecular-weight alcohols, high-molecular-weight alcohols retain a greater percentage of their combustion energy within the hydrocarbon product [63].

Oleaginous Microorganisms *Oleaginous microorganisms* accumulate triacylglycerol (TAG) within their cells, typically as a result of environmental stress such as nutrient limitations (e.g., nitrogen, phosphorous). As with algae and oil crops, the TAG can be extracted from the cell and converted to methyl esters (biodiesel) or hydrotreated to hydrocarbons [64]. In some cases, the TAG floats to the top of the fermentor and can be removed by skimming.

Hydrocarbons from Sugar (d → A) Under appropriate fermentation conditions, naturally occurring microorganisms accumulate TAG. For example, the yeast *Rhodosporidium toruloides* ferments sugar and accumulates 67% of its cellular dry weight as lipids and can produce TAG concentrations over 70 g/L in the fermentor [64]. To enhance TAG production, researchers are applying genetic engineering techniques to *Saccharomyces cerevisiae* and *Escherichia coli* [65].

Hydrocarbons from Lignocellulose (1 → A) Some microorganisms that grow directly on lignocellulose accumulate TAG within their cells [66].

Hydrocarbons from Carboxylates (e → A) Some microorganisms (e.g., *Alcanivorax borkumensis* SK2) accumulate and secrete TAG when growing on acetate [67].

Future Directions

For biofuels to make a significant impact on transportation, it is necessary to minimize land area requirements, which is accomplished by (1) growing the most productive crops and (2) utilizing the whole plant.

Crop Productivity

Among oil-producing land crops, Chinese tallow is one of the most productive and it can grow on marginally productive lands. Despite these advantages, it is not a significant source of biodiesel. In the United States, soybean oil dominates, yet it is among the least productive oil-seed crops. Chinese tallow is not widely grown because it is an invasive species and governments restrict its growth; however, the very properties that make it invasive (rapid growth, robust) make it an ideal biomass source. Perhaps plant breeders can create varieties that maintain the desirable traits, yet eliminate its invasive properties. For example, it might be possible to create varieties in which the seeds are sterile.

Among aquatic plants, water hyacinth has the highest productivity and rivals algae, which are the most productive photosynthetic organisms. Water hyacinth is also an invasive species; therefore, production systems must be developed to prevent its release into local rivers and lakes.

Among land crops, lignocellulose is the most productive. Further, numerous lignocellulosic wastes (municipal solid waste, manure, agricultural residues) are inexpensive and many are already collected.

Whole Plant Utilization

In recent years, much attention has been paid to algae because they do not compete with agricultural land, and they produce high concentrations of TAG, which is readily converted to biodiesel or hydrocarbons. Nonetheless, only a portion of the algae is TAG. Rather than focusing only on TAG, greater yields will be realized by converting the entire organism to fuel.

Given the diverse composition and high water content of algae, the best process candidate is the carboxylate platform.

Lignocellulose conversion processes use the whole plant, not just the seed or the root. Although many lignocellulose conversion processes are being developed, the technology is still immature so no processes have been fully commercialized.

Closing Thoughts

Although biofuels currently contribute to the global supply of transportation fuels, the industry is in its infancy. Once lignocellulose conversion technology is perfected, the full potential of biofuels will be realized.

Bibliography

Primary Literature

1. Renewable Fuels Association. http://www.ethanolrfa.org/pages/statistics#A
2. Energy Information Administration. http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MGFUPUS2&f=A
3. Renewable Fuels Association (2010) Ethanol industry outlook: climate of opportunity. http://www.ethanolrfa.org/page/-/objects/pdf/outlook/RFAoutlook2010_fin.pdf?nocdn=1
4. Energy Information Administration. http://www.eia.doe.gov/cneaf/solar.renewables/page/biodiesel/biodiesel.html
5. Energy Information Administration. http://www.eia.gov/dnav/pet/pet_cons_psup_dc_nus_mbblpd_a.htm
6. Zhang RD, He H, Shi XY, Zhang CB, He BQ, Wang JX (2004) Preparation and emission characteristics of ethanol-diesel fuel blends. J Environ Sci (China) 16:793–796
7. Subbaiah GV, Gopal KR, Hussain SA, Prasad BD, Reddy KT (2010) Rice bran oil biodiesel as an additive in diesel-ethanol blend for diesel engines. Int J Res Rev Appl Sci 3:334–342
8. Perlack RD, Wright LL, Turhollow AF, Graham RL, Stokes BJ, Erbach DC (2005) Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply. Oak Ridge National Laboratory, Oak Ridge. DOE/GO-102995–2135, ORNL/TM-2005/66
9. US Environmental Protection Agency (2010) Municipal solid waste generation, recycling, and disposal in the United States: facts and figures for 2009. Franklin Associates, EPA530-R-10–012. http://www.epa.gov/epawaste/nonhaz/municipal/pubs/msw2009-fs.pdf
10. Tilman D, Hill J, Lehman C (2006) Carbon-negative biofuels from low-input high-diversity grassland biomass. Science 314:1598–1600
11. Schmer MR, Vogel KP, Mitchell RB, Perrin RK (2008) Net energy of cellulosic ethanol from switchgrass. Proc Nat Acad Sci USA 105:464–469
12. McKendry P (2002) Energy production from biomass (part 1): overview of biomass. Bioresour Technol 83:37–46
13. McCollum III T, McCuistion K, Brent B (2005) Brown mid-rib and photoperiod-sensitive forage sorghums. AREC 05–20. Agricultural Program, Texas A&M University, Amarillo
14. Waclawovsky AJ, Sato PM, Lembke CG, Moore PH, Souza GM (2010) Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. Plant Biotechnol J 8:263–276
15. Westlake DF (1963) Comparisons of plant productivity. Biol Rev 38:385–425
16. Alexander AG (1985) The energy cane alternative, vol 6, Sugar series. Elsevier, Amsterdam
17. Klass DL (1998) Biomass for renewable energy, fuels, and chemicals. Academic, San Diego, 70
18. Spencer W, Bowes G (1986) Photosynthesis and growth of water hyacinth under CO_2 enrichment. Plant Physiol 82:528–533
19. http://waterquality.montana.edu/docs/irrigation/sugarbeet101.shtml
20. Freeman KC, Broadhead DM, Zummo N, Westbrook FE (1986) Sweet sorghum culture and syrup production. U.S. Department of Agriculture, Washington, DC, Handbook No. 611, 55 pp
21. Reddy BVS, Ramesh S, Reddy PS, Ramaiah B, Salimath PM, Kachapur R (2005) Sweet sorghum – a potential alternate raw material for bio-ethanol and bioenergy. SAT eJournal 1(1). www.ejournal.icrisat.org
22. US Department of Agriculture, National Agricultural Statistics Service. http://www.nass.usda.gov/
23. Tickell J (2000) From the fryer to the fuel tank: the complete guide to using vegetable oil as an alternative fuel, 3rd edn. Tickell Energy Consulting, Tallahassee, 162
24. Breitenbeck GA (2008) Chinese tallow trees a potential bioenergy crop for Louisiana. Louisiana Agr 51:10–12
25. Mata TM, Martins AA, Caetano NS (2010) Microalgae for biodiesel production and other applications: a review. Renew Sust Energy Rev 14:217–232
26. Griffiths MJ, Harrison STL (2009) Lipid productivity as a key characteristic for choosing algal species for biodiesel production. J Appl Phycol 21:493–507
27. Piskorz J, Scott DS, Radlein D (1988) Composition of oils obtained by fast pyrolysis of different woods. In: Soltes EJ, Milne TA (eds) Pyrolysis oils from biomass, vol 376, ACS symposium series. American Chemical Society, Washington, DC, pp 167–178
28. Emsley AM, Stevens GC (1994) Kinetics and mechanisms of the low-temperature degradation of cellulose. Cellulose 1:26–56
29. Klinke HB, Thomsen AB, Ahring BK (2004) Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. Appl Microbiol Biotechnol 66:10–26

30. Wu X, McLaren J, Madl R, Wang D (2010) Biofuels from lignocellulosic biomass. In: Singh OV, Harvey SP (eds) Sustainable biotechnology: sources of renewable energy. Springer, New York, pp 19–41
31. Reed TB (1981) Biomass gasification: principles and technology. Noyes Data Corporation, Park Ridge
32. Woo HC, Park KY, Kim YG, Nam I-S, Chung JS, Lee JS (1991) Mixed alcohol synthesis from carbon monoxide and dihydrogen over potassium-promoted molybdenum carbide catalysts. Appl Catal 75:267–280
33. Xu M, Lunsford JH, Goodman DW, Bhattacharyya A (1997) Synthesis of dimethyl ether (DME) from methanol over solid-acid catalysts. Appl Catal A Gen 149:289–301
34. Semelsberger TA, Borup RL, Greene HL (2006) Dimethyl ether (DME) as an alternative fuel. J Power Source 156:497–511
35. Bridgwater AV, Meier D, Radlein D (1999) An overview of fast pyrolysis of biomass. Org Geochem 30:1479–1493
36. Czernik S, Bridgwater AV (2004) Overview of applications of biomass fast pyrolysis oil. Energy Fuels 18:590–598
37. Bozell JJ, Moens L, Elliott DC, Wang Y, Neuenscwander GG, Fitzpatrick SW, Bilski RJ, Jarnefeld JL (2000) Production of levulinic acid and use as a platform chemical for derived products. Resour Conserv Recycl 28:227–239
38. Balat M, Balat H (2010) Progress in biodiesel processing. Appl Energy 87:1815–1835
39. Chisti Y (2007) Biodiesel from microalgae. Biotechnol Adv 25:294–306
40. Rein PW (1995) A comparison of cane diffusion and milling. In: Proceedings of the South African sugar technologists' association, pp 196–200
41. Granda CB, Holtzapple MT (2006) Low-pressure sugar extraction with screw-press conveyors. Int Sugar J 108:555–568
42. Wheals AE, Basso LC, Alves DMG, Amorim HV (1999) Fuel ethanol after 25 years. Trends Biotechnol 17:482–487
43. Kwiatkowski JR, McAloon AJ, Taylor F, Johnston DB (2006) Modeling the process and costs of fuel ethanol production by the corn dry-grind process. Ind Crop Prod 23:288–296
44. Ramirez EC, Johnston DB, McAloon AJ, Yee W, Singh V (2008) Engineering process and cost model for a conventional corn wet milling facility. Ind Crops Prod 27:91–97
45. Manzer LE (2010) Recent developments in the conversion of biomass to renewable fuels and chemicals. Topics Catal 53:1193–1196
46. Sierra R, Smith A, Granda C, Holtzapple MT (2008) Producing fuels and chemicals from lignocellulosic biomass. Chem Eng Prog 104:S10–S18
47. Sun Y, Cheng J (2002) Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresour Technol 83:1–11
48. Zaldivar J, Nielsen J, Olsson L (2001) Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. Appl Microbiol Biotechnol 56:17–34
49. Lynd LR, van Zyl WH, McBride JE, Laser M (2005) Consolidated bioprocessing of cellulosic biomass: an update. Curr Opin Biotechnol 16:577–583
50. Warnick TA, Methe BA, Leschine SB (2002) *Clostridium phytofermentans* sp. nov., a cellulolytic mesophile from forest soil. Int J Syst Evol Microbiol 52:1155–1160
51. Lee YY, Iyer P, Torget RW (1999) Dilute-acid hydrolysis of lignocellulosic biomass. Adv Biochem Eng Biotechnol 65:93–114
52. Lindsey TC (2010) Conversion of existing dry – mill ethanol operations to biorefineries. In: Blaschek HP, Ezeji TC, Scheffran J (eds) Biofuels from agricultural wastes and byproducts. Wiley-Blackwell, Oxford
53. Agler MT, Wrenn BA, Zinder SH, Angenent LT (2011) Waste to bioproduct conversion with undefined mixed cultures: the carboxylate platform. Trends Biotechnol 29:70–78
54. Chang HN, Kim N-J, Kang J, Jeong CM (2010) Biomass-derived volatile fatty acid platform for fuels and chemicals. Biotechnol Bioprocess Eng 15:1–10
55. Verser D, Eggeman T (2003) Process for producing ethanol. US Patent 6,509,180 B1
56. Pham V, Holtzapple M, El-Halwagi M (2010) Techno-economic analysis of biomass to fuel conversion via the MixAlco process. J Ind Microbiol Biotechnol 37:1157–1168
57. Granda CB, Holtzapple MT, Luce G, Searcy K, Mamrosh DL (2009) Carboxylate platform: the MixAlco process. Part 2: process economics. Appl Biochem Biotechnol 156:537–554
58. Datar RP, Shenkman RM, Cateni BG, Huhnke RL, Lewis RS (2004) Fermentation of biomass-generated producer gas to ethanol. Biotechnol Bioeng 86:587–594
59. Younesi H, Najafpour G, Mohamed AR (2005) Ethanol and acetate production from synthesis gas via fermentation processes using anaerobic bacterium, *Clostridium ljungdahlii*. Biochem Eng J 27:110–119
60. Vega JL, Prieto S, Elmore BB, Clausen EC, Gaddy JL (1989) The biological production of ethanol from synthesis gas. Appl Biochem Biotechnol 20(21):781–797
61. Santosh Y, Sreekrishnan TR, Kohli S, Rana V (2004) Enhancement of biogas production from solid substrates using different techniques – a review. Bioresour Technol 95:1–10
62. Oudejans JC, Van den Oosterkamp PF, Van Bekkum H (1982) Conversion of ethanol over zeolite H-ZSM-5 in the presence of water. Appl Catal 3:109–115
63. Holtzapple MT, Granda CB (2009) Carboxylate platform: the MixAlco process. Part 1: comparison of three biomass conversion platforms. Appl Biochem Biotechnol 156:525–536
64. Li Q, Du W, Liu D (2008) Perspectives of microbial oils for biodiesel production. Appl Microbiol Biotechnol 80:749–756
65. Rude MA, Schirmer A (2009) New microbial fuels: a biotech perspective. Curr Opin Microbiol 12:274–281
66. Hui L, Wan C, Hai-tao D, Xue-jiao C, Qi-fa Z, Zhao Yu-hua (2010) Direct microbial conversion of wheat straw into lipid by a cellulolytic fungus of *Aspergillus oryzae* A-4 in solid-state fermentation. Bioresour Technol 101:7556–7562
67. Manilla-Pérez E, Lange AB, Hetzler S, Steinbüchel A (2010) Occurrence, production, and export of lipophilic compounds by hydrocarbonoclastic marine bacteria and their potential use to produce bulk chemicals from hydrocarbons. Appl Microbiol Biotechnol 86:1693–1706

Books and Reviews

Bungay HR (1981) Energy, the biomass options. Wiley-Interscience, New York

Cheremisinoff NP, Cheremisinoff PN, Ellerbusch F (1980) Biomass: applications, technology, and production, vol 5, Energy, power, and environment. Marcel Dekker, New York

Klass DL (1981) Biomass as a nonfossil fuel source, vol 144, ACS symposium series. American Chemical Society, Washington, DC

Klass DL (1998) Biomass for renewable energy, fuels, and chemicals. Academic, San Diego

Moo-Young M, Blanch HW, Drew S, Wang DIC (1985) Comprehensive biotechnology, vol 3, The practice of biotechnology: current commodity products. Pergamon, New York

Saha BC, Woodward J (1997) Fuels and chemicals from biomass, vol 666, ACS symposium series. American Chemical Society, Washington, DC

Sarkanen KV, Tillman DA (1979) Progress in biomass conversion, vol 1. Academic, New York

Smil V (1983) Biomass energies: resources, links, constraints. Plenum, New York

Stafford DA, Hawkes DL, Horton R (1980) Methane production from waste organic matter. CRC Press, Boca Raton

Waldron K (2010) Bioalcohol production: biochemical conversion of lignocellulosic biomass, vol 3, Woodhead publishing series in energy. CRC Press, Boca Raton

Zaborsky OR, McClure TA, Lipinsky ES (1981) Handbook of biosolar resources, vol II, Resource materials. CRC Press, Boca Raton

Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for

Lili Huang, Matthew Barth
Department of Electrical Engineering, University of California, Riverside, CA, USA

Article Outline

Glossary

Calibration Comparison and alignment between two or more measurements or coordinates; coordinate systems are defined for different devices and the relationships between systems are established through calibration.

Computer vision Technology using computers to extract information from an image, a series of images, or video sequences to aid sensing applications.

Intelligent transportation system (ITS) System of technologies including information processing, control, communications, and system management applied to vehicles and infrastructure to improve transportation. ITS aims to improve safety, transportation efficiency, and reduce environmental impacts.

Laser range finder (LIDAR) A device that uses laser beams to determine the distance to an object. The most common technique used for distance measurement works on the time-of-flight principle.

Vehicle detection Detect, count, and classify vehicles using video camera, loop sensors, or wireless sensors. Vehicle detection systems are typically "nonintrusive" to traffic.

Vehicle tracking A combination of techniques to track a vehicle's location, record position data, and deliver data to an owner or a third party. Devices used in vehicle tracking include GPS, and possibly a video camera and other electronic devices.

Definition of the Subject

The demand on today's transportation systems is growing quite rapidly, with an estimated 30% travel demand increase predicted over the next decade [1]. Transportation infrastructure growth is not keeping pace with this traffic demand, therefore researchers and practitioners have turned to intelligent transportation systems (ITS) to improve overall traffic efficiency, thereby maximizing the current infrastructure's capacity.

ITS uses sensors, communication, and traffic control technologies to better handle the increased

demands in traffic, to enhance public safety, and to reduce environmental impacts of transportation [2]. One key ITS function that is important for many applications is the ability to gather traffic information using vehicle detection and surveillance techniques. Vehicle detection technology is widely used to provide information for vehicle counting, classification, and traffic characterization. Further, when it is implemented on a moving vehicle, it can even be utilized for vehicle navigation purposes [3]. The generalized vehicle detection problem from a moving vehicle is challenging, which aims to determine the surrounding vehicle's (relative) position, speed, and trajectory. A driver is able to determine a short-term and long-term trajectory based on the vehicle's current position and information about the surrounding vehicles.

In the case where vehicle detection is carried out on a moving vehicle, the surrounding vehicles' information is commonly collected by various sensing systems. These sensing systems typically consist a suite of sensors that can provide real-time measurements, and play an important role in the development of driver assistant systems (DAS).

One of the most common sensing techniques used is computer vision. Computer vision sensors provide a large amount of information on the surrounding environment. However, the computer vision sensors often suffer from intensity variations, narrow fields of view, and low-accuracy depth information [4]. In contrast, a laser ranging method (i.e., LIDAR) measures distance and relative angle from the sensor to the target by calculating time of flight of the laser. Its measurements depend on the size and reflectivity of the target, so the probability of detection decreases with distance [5]. Since their characteristics of computer vision and LIDAR complement each other, it is useful to integrate both computer vision and LIDAR for detecting different objects around the vehicle's environment.

A tightly coupled LIDAR and computer vision system is proposed in this entry to solve the vehicle detection problem. This sensing system is mounted on a test vehicle, as is shown in Fig. 1. A pair of LIDAR sensors is mounted on the front bumper of the vehicle, and a camera is mounted behind the front windshield.

Since sensor fusion systems are commonly used to integrate sensory data from disparate sources, the output will be more accurate and complete in comparison to the output of a single sensor. In order to effectively extract and integrate 3D information from both computer vision and LIDAR systems, the relative position and orientation between these two sensor modalities should be obtained. A sensor calibration process is used to identify the parameters that describe the relative geometric transformation between sensors [6, 7], which is a key step in the sensor fusion process. However, current calibration methods work only for visible beam LIDAR, 3D LIDAR, and 2D LIDAR [8–12]. To date, there does not exist any convenient calibration methods for multi-planar "invisible-beam" LIDAR and computer vision systems.

Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 1
The mobile sensing platform. This probe vehicle carries one camera and two IBEO ALASCA XT LIDAR sensors

A novel calibration approach of a camera with a multi-planar LIDAR system is proposed in this entry, where the laser beams are invisible to the camera. The camera and LIDAR are required to observe a planar pattern at different positions and orientations. Geometric constraints of the "views" from the LIDAR and camera images are resolved as the coordinate transformation coefficients. The proposed approach consists of two stages: solving a closed-form equation, followed by applying a nonlinear algorithm based on a maximum likelihood criterion. Compared with the classical methods which use "beam-visible" cameras or 3D LIDAR systems, this approach is easy to implement at low cost.

The combination of LIDAR and camera is employed in vehicle detection since the geometric transformation between two sensors is known from calibration. In the sensor fusion system, LIDAR sensor estimates possible vehicle positions. This information is transformed into the image coordinates. Different regions of interest (ROIs) in the imagery are defined based on the LIDAR object hypotheses. An Adaboost object classifier is then utilized to classify the vehicle in ROIs. A classifier error correction approach chooses an optimal position of the detected vehicle. Finally, the vehicle's position and dimensions are derived from both the LIDAR and image data. This sensor fusion system can be used in ITS applications such as traffic surveillance and roadway navigation tasks.

This entry is organized as follows: section "Introduction and Background" reviews background and related work, including the calibration methods for LIDAR and camera, and sensor fusion–based vehicle detection algorithms. Section "A Novel Multi-Planar LIDAR and Computer Vision Calibration Procedure Using 2D Patterns" focuses on the calibration of LIDAR and camera system. The coordinate systems of the LIDAR and camera sensors are introduced, followed by the mathematical derivation of geometric relations between the two sensors. The equations are then solved in two stages: a closed-form solution, followed by applying a nonlinear algorithm based on a maximum likelihood criterion. Section "Tightly Coupled LIDAR and Computer Vision Integrated System for Vehicle Detection" describes the sensor fusion–based vehicle detection system. Both hardware and data processing software of the sensor fusion system are introduced in this section. Finally, future work is discussed in section "Conclusions and Future Work".

Introduction and Background

During the past decade, a variety of research has been carried out in the traffic surveillance area, where numerous techniques have been developed to obtain parameters such as vehicle counts, location, speed, trajectories and classification data, for both in-vehicle navigation and freeway traffic surveillance applications [13].

As one of the most popular traffic surveillance techniques, computer vision-based approaches are one of the most widely used and promising techniques. LIDAR is another attractive technology due to its high accuracy in ranging, wide-area field of view, and low data processing requirements [5]. The other sensors used in vehicle detection include radar and embedded loop sensors. A brief comparison of the sensor technologies and their advantages as well as disadvantages is given in Table 1.

Sensor fusion systems in the vehicle detection application aim to gather information from the far-field as well as near-field sensors, and combine them in a meaningful way [15]. The output of the sensor fusion system should be the states of objects around the test vehicle.

In this section, a few of the most popular LIDAR sensors that are available commercially today are introduced. This is followed by a discussion of current LIDAR and computer vision calibration methods. The calibration methods include the visible LIDAR beam–based calibration, 3D LIDAR and the 2D single planar LIDAR calibration. Sensor fusion techniques for vehicle detection and tracking systems are also discussed here.

Laser Range Finder (LIDAR)

The vehicle detection solution aims to estimate the states of surrounding vehicles. Vehicle state includes position, orientation, speed, and acceleration. State estimation addresses the problems of estimating quantities from sensors that are not directly observable [16].

LIDAR sensors are commonly utilized in vehicle navigation for detecting surrounding vehicles, infrastructure, and pedestrians. It can also be used in vehicle

Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Table 1 Performance comparison of existing sensor technologies used in ITS [14]

Sensor technology	Advantage	Disadvantage
LIDAR	Detect distance and angle with high accuracy	High cost
	Low data processing requirement	Limited detection range
	Operational in fog and rain	Difficult to classify the object
Radar	Direct measurement of speed or distance	Relatively low precision
	Compact size	Limited field of view
	Low data processing requirement	May have identification problem in multilane applications
		Medium cost
Video camera	Provide real-time image of traffic	High requirement for data processing and storage
	Low cost	
	Multiple lanes observed	Different algorithms required for day- and nighttime
	No traffic interruption for installation and repair	
	Large field of view	Susceptible to atmospheric obscurants and weather change
Infrared camera	Day and night operation	High requirement for data processing
	Operational in fog	Susceptible to weather change
Inductive loop detector	Low cost per unit	Installation and maintenance require traffic disruption
	Large experience base	Easily damaged

localization, either as the only sensor or in some combination with GPS and INS.

One of the most popular LIDAR sensors is the SICK LMS 2xx series [17]. A SICK LIDAR operates at distance up to 80 m with an angular resolution of 0.5° and a measurement accuracy of typically 5 cm. The distance between the sensor and an object is calculated by measuring the time interval between an emitted laser pulse and a reception of the reflected pulse. Amplitude of the received signal is used to determine reflectivity of the object surface. The SICK LIDAR is able to detect dark objects at long ranges. Moreover, compared to the CCD cameras and RADAR systems, the view angle of SICK LIDAR is larger, e.g., 180°. Figure 2a illustrates the SICK LMS200 LIDAR.

The HOKUYO UXM-30LN LIDAR is another single planar range sensor for intelligent robots and vehicles [18]. Its detection range is up to 60 m, and the horizontal field of view is 190°. The distance accuracy is 30 mm when the range is less than 10 m, and 50 mm when the range is between 10 and 30 m. The angular resolution is 0.25°. The device is shown in Fig. 2b.

As another example, the ALASCA XT laser scanner made by IBEO is a multi-planar LIDAR, which splits the laser beam into four vertical layers. The aperture angle is 3.2°. The distance range is up to 200 m, and the horizontal field of view is 240° [19]. Figure 2c shows the IBEO sensor.

The Velodyne HDL-64E LIDAR is a 3D sensor which is specifically designed for autonomous vehicle navigation [20]. With 360° horizontal by 25° vertical field of view, 0.09° angular resolution, and 10 Hz refresh rate, Velodyne provides surrounding 3D traffic information with high accuracy (<5 cm resolution)

Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 2
A variety of LIDAR sensors: (**a**) SICK LMS LIDAR, (**b**) HOKUYO UXM-30LN LIDAR, (**c**) IBEO ALASCA XT LIDAR, and (**d**) Velodyne HDL-64E LIDAR

and efficiency. The detection range is 100 m for cars, and the latency is less than 0.05 ms. Figure 2d illustrates the Velodyne sensor.

LIDAR sensors were initially used by automatic guided vehicles in indoor environments. More recently, the performance of these ranging sensors has been improved, so that they now can be used in outdoor environments on vehicles. A primary application example is the DARPA Urban Challenge in 2007, in which autonomous vehicles were capable of driving in traffic and performing complex maneuvers such as acceleration and deceleration, lane change, and parking [13]. IBEO and SICK LIDARs were used in many of the finalists for object detection and localization. The Velodyne sensor was used by the five out of six of the finishing teams.

LIDAR and Computer Vision Calibration

Sensor fusion systems are commonly used to combine the sensory data from disparate sources, so that the result will be more accurate and complete in comparison to the output of one sensor. As an example, the winner of the 2007 DARPA Urban Grand Challenge performed sensor fusion between a GPS receiver, long- and short-range LIDAR sensors, and stereo cameras [21].

In order to effectively extract and integrate 3D information from both computer vision and LIDAR systems, the relative position and orientation between these two sensor modalities can be obtained. The relative geometric transformation can be solved through a calibration process [6, 7]. Several approaches have been defined and utilized for LIDAR and computer vision calibration. These techniques can be roughly classified into three categories.

Visible Beam Calibration Visible beam calibration is performed by using cameras to observe the LIDAR beams or reflection points. The calibration system usually consists of an active LIDAR and some infrared or near-infrared cameras. The LIDAR system typically projects stripes with a known frequency, while these stripes are visible to the camera [8–10]. For example, the LIDAR beams used in [9] are captured by a 955fps high-speed camera. However, the image of the high-speed camera is not suitable for monitoring. Color image of the LIDAR beams is generated by letting the vision output go through a beam splitter.

This approach requires a high-cost infrared camera, which should be sensitive to the spectral emission band of the LIDAR. Therefore, this method is not suitable for the low-cost sensor fusion systems.

Three-Dimensional (3D) LIDAR-Based Calibration The 3D LIDAR-based calibration method calibrates the computer vision system with a 3D LIDAR system. Various features are captured by both the camera and the LIDAR. These features are usually in the form of planes, corner, or edges of specific calibration object. An elaborate setup is required. Moreover, dense LIDAR beams in both the vertical and horizontal directions are necessary for the calibration.

The 3D calibration algorithm presented in [11] uses checkerboard in calibration, which is commonly used

for camera calibration. Coefficients of the checkerboard plane are first calculated by LIDAR, and then the coefficients are computed by camera in computer vision coordinates. A two-stage optimization procedure is implemented to minimize the distance between the calculated results and the measurement output.

When the features are edges or corners, accuracy of the calibration method depends on the accuracy by which features are localized [22]. When the features are planes, the LIDAR beams must be sufficiently dense [11]. Therefore, these methods cannot easily be applied to single planar or sparse multi-planar LIDAR systems.

Two-Dimensional (2D) Planar-Based Calibration This approach works for the calibration of camera and 2D LIDAR integration system. The calibration system proposed in [23] consists of a monochrome CCD camera and a LIDAR. The camera and the LIDAR have been pre-calibrated so that their coordinates are parallel to each other. A "V"-shaped pattern is designed to obtain the translation between these two sensors. The calibration procedure is implemented in two steps: the LIDAR detects the "V" shape and finds the vertex, and camera detects the intersection line which cuts the pattern into two parts.

Another calibration approach is proposed in [12] using a checkerboard for calibration. This method is based on observing a plane of an object and solving distance constraints from the camera and LIDAR systems. This approach works for a single planar LIDAR only, e.g., the SICK LMS 2xx series LIDAR.

To date, there does not exist any convenient calibration methods for multi-planar "invisible-beam" LIDAR and computer vision systems. Section "A Novel Multi-Planar LIDAR and Computer Vision Calibration Procedure Using 2D Patterns" proposes a method to handle this case, which is the first calibration method for this system as to the author's best knowledge.

Sensor Fusion–Based Vehicle Detection and Tracking

Computer vision is generally used on mobile platform–based object detection and tracking systems, either separately or along with LIDAR sensors [24]. Most of the computer vision techniques utilize a simple segmentation method such as background subtraction or temporal difference to detect objects [25]. However, these approaches suffer with the fast background changes due to camera motion. A trainable object detection method is proposed in [26] based on a wavelet template, which defines the shape of an object in terms of a subset of the wavelet coefficients of the image. However, the application of vision sensors in vehicle navigation is far from sufficient: clustering, illumination, occlusion, among many other factors, affect the overall performance. Fusion of camera and active sensors such as LIDAR or RADAR, is being investigated in the context of on-board vehicle detection and classification.

A LIDAR and a monocular camera–based detection and classification system is proposed in [27]. Detection and tracking are implemented in the LIDAR space, and the object classification work both in LIDAR space (Gaussian Mixture Model classifier) and in computer vision system (Adaboost classifier). A Bayesian decision rule is proposed to combine the results from both classifiers, and thus a more reliable classification is achieved.

Another integration structure is proposed in [28], in which a LIDAR is integrated with a far-infrared camera and an ego motion sensor. LIDAR-based shape extraction is employed to select region of interests (ROIs). This system combines a straightforward methodology with a backward loop one. Kalman filtering is used as the data fusion algorithm.

A similar technique is presented in [29], which makes use of RADAR, velocity, and steering sensors to generate position hypotheses. Examination of the hypotheses is implemented by a computer vision sensor. Classification is performed using a shape model for either the monocular camera vision or the infrared spectrum images.

A Novel Multi-planar LIDAR and Computer Vision Calibration Procedure Using 2D Patterns

In this section, a novel calibration approach is proposed for a LIDAR and computer vision sensor fusion system. This system consists of a camera with a multi-planar LIDAR, where the laser beams are invisible to the camera. This calibration method also works for computer vision and 3D LIDAR systems.

Although several calibration methods have been developed to obtain the geometric relationship

between two sensors, only a few of them have provided a complete sensitivity analysis of the calibration procedure (see section "LIDAR and Computer Vision Calibration"). As part of the calibration method proposed in this section, the effects of LIDAR noise level as well as total number of poses on calibration accuracy are also discussed.

This section is organized as follows: section "Sensor Alignment" gives the setup using planar planes and defines the calibration constraint. Section "Calibration Solutions" describes in detail how to solve this constraint in two steps. Both a closed-form solution and a nonlinear minimization solution based on maximum likelihood criterion are introduced. Experimental results with different poses are provided in section "Experimental Results". Finally, a brief summary is given in section "Summary".

Sensor Alignment

The setup for a multi-planar LIDAR and camera calibration is described here.

Sensor Configuration In the calibration system, an instrumented vehicle is equipped with two IBEO ALASCA XT LIDAR sensors which are mounted on the front bumper. The LIDAR sensor scans with four separate planes. The distance range is up to 200 m, the horizontal field of view angle of a single LIDAR is 240°, and the total vertical field of view for the four planes is 3.2°. A camera is mounted on the vehicle behind the front windshield, as is shown in Fig. 1.

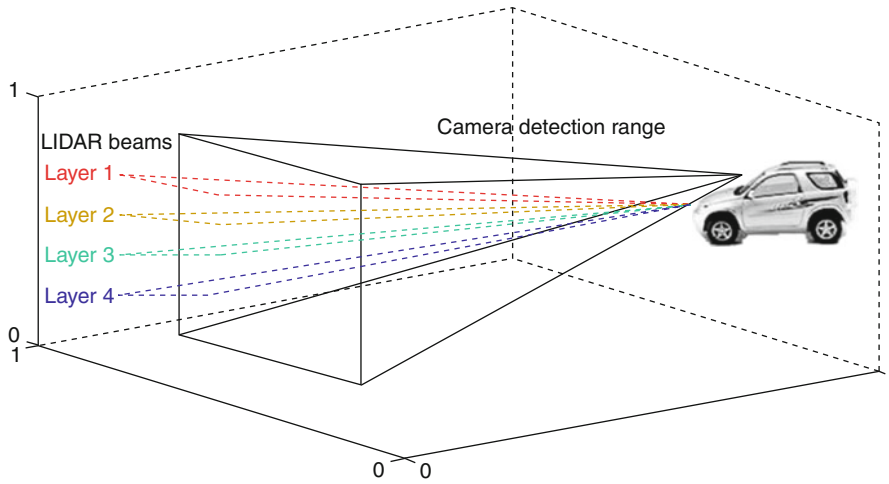
In order to use the measurements from different kinds of sensors at various positions on the vehicle, the measurements should be transformed from their own coordinate into some common coordinate system. This section focuses on obtaining the spatial relationship between video and LIDAR sensors. The geometric sensor model is shown in Fig. 3.

Vision, LIDAR, and World Coordinate Systems There are several coordinate systems in the overall system to be considered: the camera coordinates, the LIDAR coordinates, and the world coordinate systems.

A camera can be represented by the standard pinhole model. One 3D point in the camera coordinate denoted by $\mathbf{P}_c = [\mathbf{X}_c \quad \mathbf{Y}_c \quad \mathbf{Z}_c]^T$ is projected to a pixel $\mathbf{p} = [u \quad v]^T$ in the image coordinate. The pinhole model is given as [30]:

$$s\mathbf{p} \sim \mathbf{A}[\mathbf{R} \quad \mathbf{t}]\mathbf{P}_c \quad \text{with} \quad \mathbf{A} = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where s is an arbitrary scale factor. $\mathbf{A}$ is the camera intrinsic matrix defined by coordinates of the principal point $(u_0 \quad v_0)$, scale factors α and β in image u and v



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 3
Geometric model with the camera and the LIDAR

axes, and skewness of the two image axes γ. ($\mathbf{R}$, $\mathbf{t}$) are called extrinsic parameters. The 3×3 orthonormal rotation matrix $\mathbf{R}$ represents the orientation of world coordinates to the camera coordinate system. The translation matrix $\mathbf{t}$ is a three-vector representing origin of world coordinates in the camera's frame of reference. In the real world, lens of the camera may also have image distortion coefficients, which include radial and tangential distortions and are usually stored in a five-vector [31]. In this entry, the lens is assumed to have no significant distortion, or the distortion has already been eliminated.

The LIDAR sensor provides distance and direction of each scan point in LIDAR coordinates. Distances and directions can be converted into a 3D point denoted by $\mathbf{P}_l = [X_l \quad Y_l \quad Z_l]^T$ [19]. The origin of LIDAR coordinates is the equipment itself. X, Y, and Z axes are defined as forward, leftward, and upward from the equipment, respectively. The camera and LIDAR reference systems are shown in Fig. 4.

In addition to the camera and LIDAR reference systems, another coordinate system is used in the calibration procedure: the world frame of reference. In the calibration process, a checkerboard is placed in front of the sensors. The first grid on the upper-left corner of this board is defined to be the origin of the world coordinates [31].

Suppose a fixed point $\mathbf{P}$ is denoted as $\mathbf{P}_c = [X_c \quad Y_c \quad Z_c]^T$ in the camera coordinates, and $\mathbf{P}_l = [X_l \quad Y_l \quad Z_l]^T$ in the LIDAR coordinates. The transformation from LIDAR coordinate to camera coordinate is given as:

$$\mathbf{P}_c = \mathbf{R}_l^c \mathbf{P}_l + \mathbf{t}_l^c \tag{2}$$

where $(\mathbf{R}_l^c, \quad \mathbf{t}_l^c)$ are the rotation and translation parameters which relate LIDAR coordinate system to the camera coordinate system.

The purpose of this calibration method is to solve Eq. 2 and obtain coefficients $(\mathbf{R}_l^c, \quad \mathbf{t}_l^c)$, so that any given point in the LIDAR reference system can be transformed to the camera coordinates.

Basic Geometric Interpretation A checkerboard visible to both sensors is used for calibration. In the following sections, the planar surface defined by the checkerboard is called the *checkerboard plane*. Without loss of generality, the checkerboard plane is assumed to be on $Z = 0$ in the world coordinates. Let $\mathbf{r}_3$ denotes the i-th column of the rotation matrix $\mathbf{R}$. $\mathbf{r}_3$ is also the surface normal vector of the calibration plane in camera coordinate systems [31].

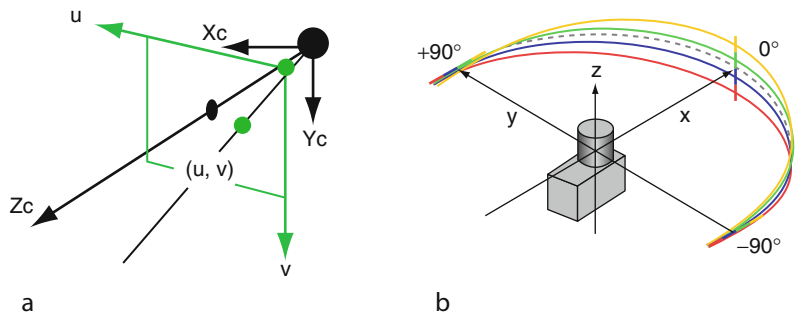
Note the origin of world coordinate is the upper-left corner of the checkerboard, and the origin of camera coordinate is the camera itself. The translation vector $\mathbf{t}$ represents relative position of the checkerboard's upper-left corner in the camera's reference system. Since both $\mathbf{t}$ and $\mathbf{P}_c$ are points on the checkerboard plane denoted in camera coordinates, a vector $\vec{v}$ is defined as $\vec{v} = \mathbf{P}_c - \mathbf{t}$. Note that $\vec{v}$ is a vector on the checkerboard plane, and $\mathbf{r}_3$ is orthogonal to this plane, so:

$$\mathbf{r}_3 \cdot \vec{v} = 0 \tag{3}$$

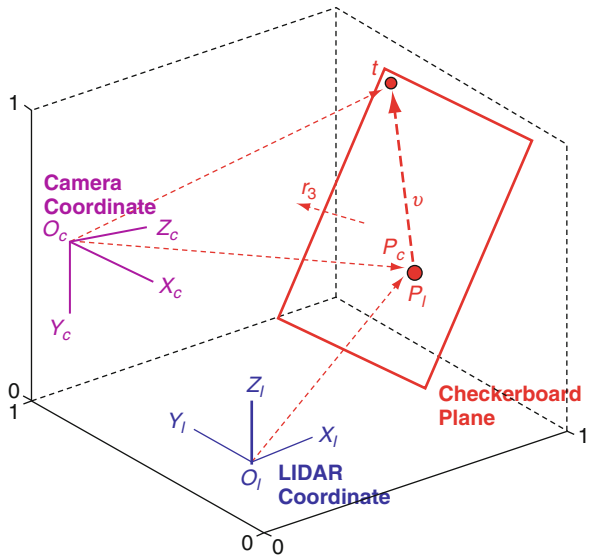
where $\cdot$ denotes the inner product. The geometric interpretation for Eq. 3 is illustrated in Fig. 5.

By substituting Eq. 2 into Eq. 3, Eq. 3 becomes:

$$\mathbf{r}_3^T \left(\mathbf{R}_l^c \mathbf{P}_l + \mathbf{t}_l^c - \mathbf{t}\right) = 0 \tag{4}$$



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 4
Two coordinate systems. (**a**) The camera coordinates and screen coordinate systems, and (**b**) the LIDAR coordinate system



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 5
Geometric interpretation of the camera coordinates, the LIDAR coordinates, and checkerboard

Since point $\mathbf{P}_l$ in LIDAR coordinates is $[X_l \quad Y_l \quad Z_l]^T$, from Eq. 4:

$$\mathbf{r}_3^T \left[\mathbf{R}_l^c \quad \mathbf{t}_l^c - \mathbf{t}\right] \begin{bmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{bmatrix} = 0 \tag{5}$$

For each LIDAR point on the checkerboard plane, Eq. 5 explains the geometric relationships and constraints on $\left(\mathbf{R}_l^c, \quad \mathbf{t}_l^c\right)$. This is the basic constraints for the calibration from the LIDAR to the vision coordinate system.

Calibration Solutions

This subsection provides the method to efficiently obtain calibration coefficients $\left(\mathbf{R}_l^c, \quad \mathbf{t}_l^c\right)$. An analytical solution is proposed, followed by a nonlinear optimization technique based on the maximum likelihood criterion.

Closed-Form Solution Initially, the camera's intrinsic parameters are calibrated using a standard Camera Calibration Toolbox [31]. For each pose of the checkerboard, there is one set of camera extrinsic parameters $(\mathbf{R}, \quad \mathbf{t})$. Each $(\mathbf{R}, \quad \mathbf{t})$ is determined also using the toolbox, after which $\mathbf{r}_3$ and $\mathbf{t}$ in Eq. 5 are obtained.

For simplicity, it is defined that $\mathbf{r}_3 = [r_{31} \quad r_{32} \quad r_{33}]^T$, $\Delta_i = \mathbf{t}_l^c - \mathbf{t} = [\Delta_x \quad \Delta_y \quad \Delta_z]^T$, and m_{ij} is the element on the i-th row, j-th column in matrix $\mathbf{R}_l^c$. Suppose for one pose of the checkerboard, there are p LIDAR points on the checkerboard plane, denoted as $\mathbf{P}_{l,1} = [X_{l,1} \quad Y_{l,1} \quad Z_{l,1}]^T$, $\mathbf{P}_{l,2} = [X_{l,2} \quad Y_{l,2} \quad Z_{l,2}]^T, \ldots, \mathbf{P}_{l,p} = [X_{l,p} \quad Y_{l,p} \quad Z_{l,p}]^T$. The geometric interpretation becomes a $\mathbf{Ax} = \mathbf{0}$ problem, where $\mathbf{A}$ is a $N \times 12$ matrix, and $\mathbf{x}$ is a 12-vector to be solved. $\mathbf{A}$ and $\mathbf{x}$ are given in Eq. 6.

$$\begin{aligned} \mathbf{A} &= [r_{31}\mathbf{P}_{l,p} \; r_{31}\mathbf{E} \; r_{32}\mathbf{P}_{l,p} \; r_{32}\mathbf{E} \; r_{33}\mathbf{E}_{l,p} \; r_{33}\mathbf{E}] \\ \mathbf{x} &= [\mathbf{m}_1 \, \Delta_x \, \mathbf{m}_2 \Delta_y \, \mathbf{m}_3 \, \Delta_z]^T \end{aligned} \tag{6}$$

where $\mathbf{P}_{l,p} = \left[\mathbf{P}_{l,1} \;\; \mathbf{P}_{l,2} \;\; \ldots \;\; \mathbf{P}_{l,p}\right]^T$, $[\mathbf{E} = [1 \;\; 1 \;\; \cdots \;\; 1]^T$ is a $(p \times 1)$ vector, and $\mathbf{m}_i = [m_{i1} \;\; m_{i2} \;\; m_{i3}]$, $i = 1, 2, 3$.

By getting the LIDAR points $\mathbf{P}_{l,1}$, $\mathbf{P}_{l,2}$, …, $\mathbf{P}_{l,p}$, $\mathbf{x}$ is estimated using the least square method. In order to avoid the solution $\mathbf{x} = \mathbf{0}$, normalization constraints are proposed. Faugeras and Toscani suggested the constraint $m_{31}^2 + m_{32}^2 + m_{33}^2 = 1$, which is singularity free [32]. This restriction is proposed from the coincidence that $[m_{31} \quad m_{32} \quad m_{33}]$ is the third row of the rotation matrix $\mathbf{R}_l^c$. Thus solving the equation $\mathbf{Ax} = \mathbf{0}$ is transformed into minimizing the norm of $\mathbf{Ax}$, i.e., minimizing $|\mathbf{Ax}|$ with the restriction $m_{31}^2 + m_{32}^2 + m_{33}^2 = 1$.

$|\mathbf{Ax}|$ can be minimized using a Lagrange method [33]. Let $\mathbf{m}_3 = [m_{31} \; m_{32} \; m_{33}]$ and $\mathbf{m}_9$ be a vector containing the remaining nine elements in $\mathbf{x}$. The Lagrange equation is written as:

$$L = \mathbf{A}_9 \cdot \mathbf{m}_9 + \mathbf{A}_3 \cdot \mathbf{m}_3 + \lambda\left(\mathbf{m}_3^T \mathbf{m}_3 - 1\right) \tag{7}$$

where $\mathbf{A}_3$ contains the 9th to 11th columns of $\mathbf{A}$, and $\mathbf{A}_9$ contains the remaining nine columns corresponding to $\mathbf{m}_9$.

The closed-form linear solution is:

$$\begin{aligned} \lambda\mathbf{m}_3 &= \left(\mathbf{A}_3^T\mathbf{A}_3 - \mathbf{A}_3^T\mathbf{A}_9\left(\mathbf{A}_9^T\mathbf{A}_9\right)^{-1}\mathbf{A}_9^T\mathbf{A}_3\right)\mathbf{m}_3 \\ \mathbf{m}_9 &= -\left(\mathbf{A}_9^T\mathbf{A}_9\right)^{-1}\mathbf{A}_9^T\mathbf{A}_3\mathbf{m}_3 \end{aligned} \tag{8}$$

It is well known that $\mathbf{m}_3$ is the eigenvector of the symmetric positive definite matrix $\mathbf{A}_3^T\mathbf{A}_3 - \mathbf{A}_3^T\mathbf{A}_9\left(\mathbf{A}_9^T\mathbf{A}_9\right)^{-1}\mathbf{A}_9^T\mathbf{A}_3$ associated with the smallest eigenvalue. $\mathbf{m}_9$ is obtained after $\mathbf{m}_3$. Once $\mathbf{m}_3$

and $\mathbf{m}_9$ are known, the rotation and translation matrix $\left(\mathbf{R}_l^c, \mathbf{t}_l^c\right)$ is available.

Because of data noise, the rotation matrix $\mathbf{R}_l^c$ may not in general satisfy $\left(\mathbf{R}_l^c\right)^T \mathbf{R}_l^c = \mathbf{I}$. One solution is to obtain $\hat{\mathbf{R}}_l^c$, which is the best approximation of given $\mathbf{R}_l^c$. This $\hat{\mathbf{R}}_l^c$ has the smallest Frobenius norm of the difference $\hat{\mathbf{R}}_l^c - \mathbf{R}_l^c$, subject to $(\hat{R}_l^c)^T \hat{\mathbf{R}}_l^c = \mathbf{I}$ [30].

Maximum Likelihood Estimation The closed-form solution is obtained by minimizing an algebraic distance $|\mathbf{Ax}|$, which is not physically meaningful. In this subsection, the problem is refined through maximum likelihood function using multi-pose checkerboard planes, which is more meaningful.

In the proposed camera calibration approach, differences of image points and the corresponding projection of the ground truth point in an image are minimized [30]. This method is also valid for visible-beam LIDAR calibration [12]. In the test, the Euclidean distances from camera to the checkerboard are checked. Note that Eq. 4 can be written as:

$$\mathbf{r}_3^T\left(\mathbf{R}_l^c \mathbf{P}_l + \mathbf{t}_l^c\right) = \mathbf{r}_3^T \mathbf{t} \tag{9}$$

where both $\mathbf{R}_l^c \mathbf{P}_l + \mathbf{t}_l^c$ and $\mathbf{t}$ are points on the calibration plane surface, and $\mathbf{r}_3$ is the normal vector to this surface. Therefore, both the left and right sides of Eq. 9 are the distance between the checkerboard plane and the origin of the camera reference system.

Suppose there are totally n poses of the calibration plane. For the i-th pose, there is a set of $(\mathbf{r}_3, \mathbf{t})$ denoted as $\left(\mathbf{r}_3^i, \mathbf{t}^i\right)$. LIDAR points are assumed to be corrupted by Gaussian distributed noise. The maximum likelihood function is defined by minimizing sum of the difference between $\mathbf{r}_3^T\left(\mathbf{R}_l^c \mathbf{P}_l + \mathbf{t}_l^c\right)$ and $\mathbf{r}_3^T \mathbf{t}$ for all the LIDAR points. Suppose for the i-th plane, there are p_i LIDAR points. The solution satisfies:

$$\arg\min_{\mathbf{R}_l^c, \mathbf{t}_l^c} \sum_{i=1}^{n} \frac{1}{p_i} \sum_{j=1}^{p_i} \left(\left(\mathbf{r}_3^i\right)^T \left(\mathbf{R}_l^c \mathbf{P}_{l,j}^i + \mathbf{t}_l^c\right) - \left(\mathbf{r}_3^i\right)^T \mathbf{t}^i \right)^2 \tag{10}$$

where $\mathbf{R}_l^{c,i} \mathbf{P}_{l,j}^i + \mathbf{t}_l^{c,i}$ is the coordinate of $\mathbf{P}_{l,j}^i$ in the camera reference system, according to Eq. 2.

By using Rodriguez formula [32], the rotation matrix $\mathbf{R}_l^c$ is transformed into a vector, which is parallel to the rotation axis and whose magnitude is equal to the rotation angle. Thus $\left(\mathbf{R}_l^c, \mathbf{t}_l^c\right)$ forms a vector. Equation 10 is solved using the Levenberg–Marquardt algorithm (LMA) [34, 35], which provides numerical solutions to the problem of minimizing nonlinear functions. LMA requires an initial guess for the parameters to be estimated. In this algorithm, $\left(\mathbf{R}_l^c, \mathbf{t}_l^c\right)$ in the closed form is used as this initial state. For each pose, a set of $\left(\mathbf{R}_l^c, \mathbf{t}_l^c\right)$ is obtained. The weighted average is used as an initial guess, where the scalar weight is normalized as a relative contribution of each checkerboard pose. Then LMA gives a robust solution even if the initial state starts far off the final solution.

Summary of Calibration Procedure The calibration procedure proposed in this approach can be summarized as:

1. Place the checkerboard in view of the camera and LIDAR systems. Make sure that the plane is within the detection zone of both sensors. The different poses of checkerboard cannot be parallel to each other, otherwise the parallel poses do not provide enough constraints on $\mathbf{R}_l^c$.
2. Take a few measurements (images) of the checkerboard under different orientations. For each orientation, read the LIDAR points on this plane from the output.
3. Estimate the coefficients using the closed-form solution given in section "Closed-Form Solution".
4. Refine all the coefficients using the maximum likelihood estimation in section "Maximum Likelihood Estimation".

Experimental Results

The proposed vision-LIDAR calibration algorithm has been tested on both a computer simulation platform and with real-world data.

Computer Simulations The camera is assumed to have been calibrated. It is simulated to have the following properties: $\alpha = 1,200$, $\beta = 1,000$, and the skewness coefficient $\gamma = 0$. The principal point is $(320, 240)$, and the image resolution is 640×480. The calibration checkerboard consists of 10×10 grids. The size of each square grid is 5cm × 5cm. The position and orientation of the LIDAR relative to the camera have also been defined. The LIDAR's position in camera coordinates is $\mathbf{t}_l^c = [\,10 \quad 150 \quad 100\,]^T$ centimeters, and the rotation

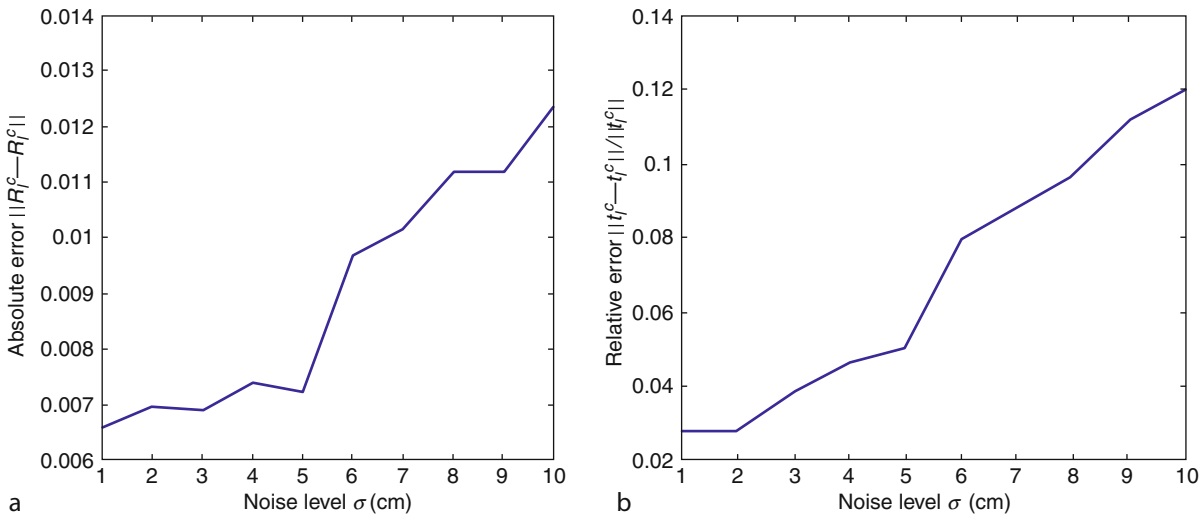
matrix $\mathbf{R}_l^c$ is parameterized by a three-vector rotation vector $[-85^o \quad 10^o \quad -80^o]^T$.

The LIDAR points are calculated based on the location of the camera, and relative position as well as orientation of the checkerboard. Gaussian noise is added to the points.

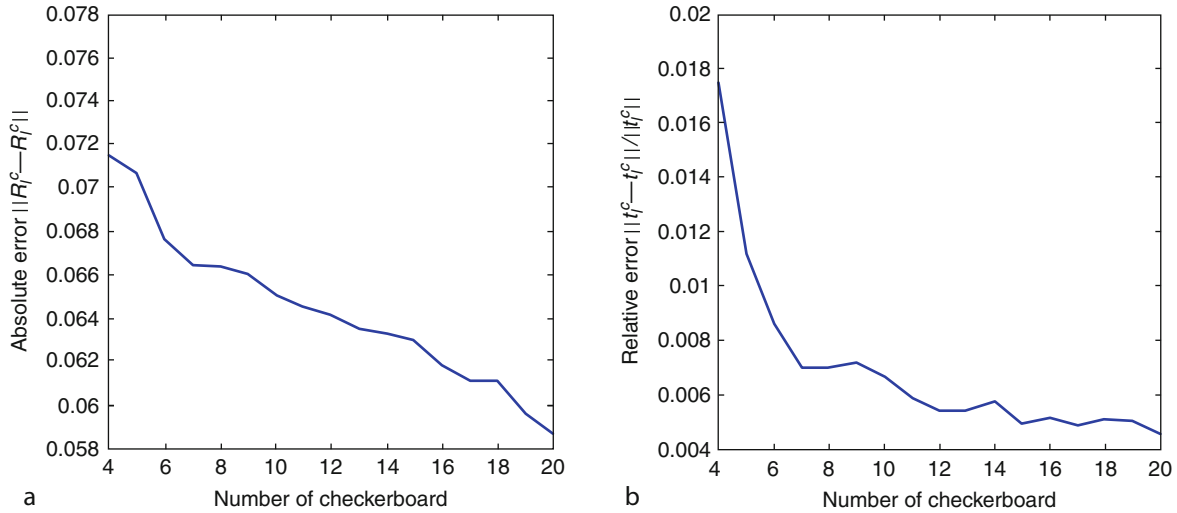
Performance with respect to Gaussian Noise Level A checkerboard plane is placed in front of the camera and the LIDAR. Three poses are used here. All of them have $\mathbf{t} = [-20 \quad -20 \quad -550]^T$. Three rotation matrices are defined by the rotation vectors as $\mathbf{r}_1 = [170^o \quad -5^o \quad 85^o]^T$, $\mathbf{r}_2 = [170^o \quad 15^o \quad 85^o]^T$, $\mathbf{r}_3 = [170^o \quad -25^o \quad 85^o]^T$, respectively. Gaussian noise with zero mean and standard deviation (from 1 to 10 cm) is added to the LIDAR points. The estimation results are then compared with ground truth. For each noise level, 100 independent random trials are delivered. The averaged calibration error is shown in Fig. 6, where the calculation results are denoted as $\hat{\mathbf{R}}_l^c$ and $\hat{\mathbf{t}}_l^c$, respectively. This figure illustrates that the calibration error increases with noise level, as expected. For $\sigma < 7$cm (which is larger than the normal standard deviation for most LIDAR sensors), the error of norm of $\mathbf{R}_l^c$ is less than 0.01. With three checkerboard poses, the relative translation error is less than 5% when $\sigma < 5$cm.

Performance with respect to the Number of Checkerboard Positions The checkerboard was originally setup as parallel to the image plane. Then it is rotated by, where the rotation axis is randomly selected in a uniform sphere. The number of checkerboards used for calibration varies from 4 to 20. Gaussian noise with zero mean and standard deviation $\sigma = 4$cm is added to the LIDAR points. For each position, 100 trials of independent rotation axes are implemented. The averaged result is illustrated in Fig. 7. This figure shows that when the number of checkerboard positions increases, the calibration error decreases.

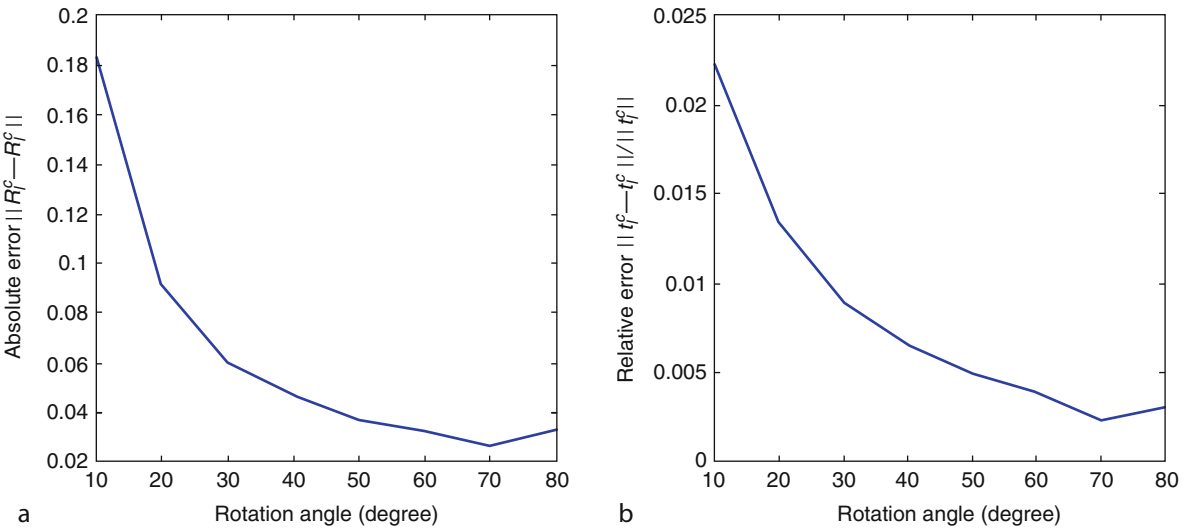
Performance with respect to the Orientation of Checkerboard The checkerboard plane is initially set as parallel to the image plane. It is then rotated around a randomly chosen axis with angle θ. The rotation axis is randomly selected from a uniform sphere. The rotation angle θ varies from 10^o to 80^o, and 10 checkerboards are used for each θ. Gaussian noise with zero mean and standard deviation $\sigma = 4$cm is added to the LIDAR points. For each rotation angle, 100 trials are repeated and the average error is calculated. The simulation result is shown in Fig. 8. The calibration error decreases when the rotation angle increases. When the rotation angle is too small, the calibration planes are almost parallel to each other,



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 6
Rotation and translation error with respect to the noise level. (**a**) Rotation error with respect to noise level. (**b**) Translation error with respect to noise level



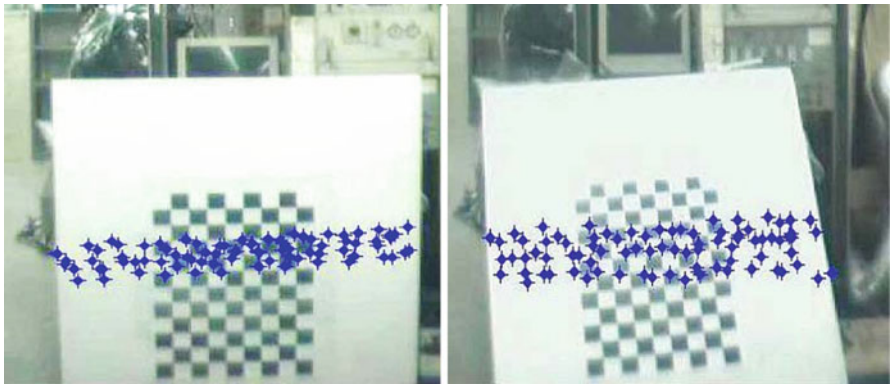
Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 7
Rotation and translation error with respect to number of checkerboard positions. (**a**) Rotation error with respect to number of checkerboard. (**b**) Translation error with respect to number of checkerboard



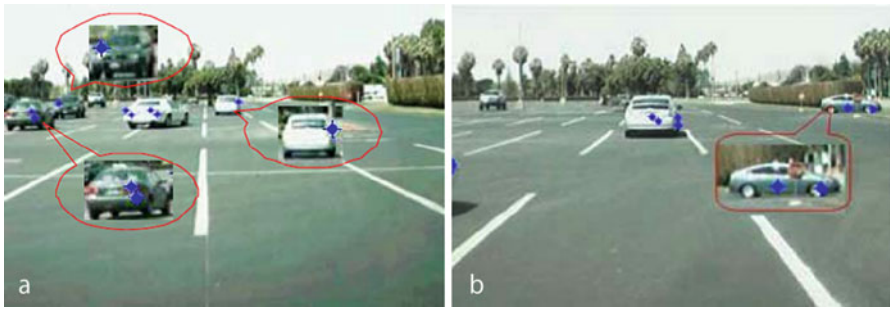
Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 8
Rotation and translation error with respect to the orientation of the checkerboard plane. (**a**) Rotation error with respect to orientation of the checkerboard plane. (**b**) Translation error with respect to orientation of the checkerboard plane

which cause error. When the rotation angle is too large, the calibration plane is almost perpendicular to the image plane, which makes the LIDAR measurement less precise.

Real Data Calibration The calibration method has been tested using an IBEO ALASCA XT LIDAR system and a Sony CCD digital camera with a 6 mm lens. The image resolution is 640 × 480. The checkerboard plane



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 9
Two checkerboard positions. The LIDAR points are indicated by *blue dots*. The calibration method proposed in this entry is used to estimate $\mathbf{R}_l^c$ and $\mathbf{t}_l^c$



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 10
Two sensor fusion image frames. The *red rectangle* is an enlarged image of the detected area

consists of a pattern of 16×16 squares, so there are totally 256 grids on the plane. The size of each grid is 2.54cm $\times$ 2.54cm (1×1 in.).

Twenty images of the plane were taken with different orientations, and the LIDAR points are recorded simultaneously. Two examples of the calibration results are shown in Fig. 9, where the LIDAR points are mapped to image reference system using estimated $\mathbf{R}_l^c$ and $\mathbf{t}_l^c$. Although the ground truth of $\mathbf{R}_l^c$ and $\mathbf{t}_l^c$ are not known, Fig. 9 shows that the estimation results are pretty reasonable.

Application in Vehicle Detection The calibration method has been integrated into a mobile sensing system. This mobile sensing system is designed to detect and track surrounding vehicles, which is the first and fundamental step for any of the automatic traffic surveillance systems. However, object detection is a big challenge for the moving platform. Both the foreground and the background are rapidly changing, which makes it difficult to extract the foreground regions from the background. The sensor fusion technique is used to compensate for the spatial motion of the moving platform. Figure 10 gives two images from the vehicle detection video.

In Fig. 10a, there are totally four vehicles detected by the LIDAR, where the farthest vehicle is 55 m away from the mobile sensing system. It is hard to obtain the vehicle's distance and orientation from an image alone. The LIDAR points provide a reliable estimation of this vehicle's position. In Fig. 10b, one car parallel to the probe vehicle is detected by the LIDAR. Meanwhile, it is

partially visible in the image, together with its shadow on the ground. Although this vehicle is hardly recognizable in the image, with a wide angle of view, LIDAR data provide enough information to reconstruct the location of this vehicle.

The experiment results illustrate that the calibration algorithm provides good results. The sensor fusion system combining LIDAR and computer vision information sources presents distance and orientation information. This system is helpful for vehicle detection and tracking applications.

Summary

In this section, a novel calibration algorithm was developed to obtain the geometry transformation between a multi-plane LIDAR system and a camera vision system. This calibration method requires LIDAR and camera to observe a checkerboard simultaneously. A few checkerboard poses are observed and recorded. The calibration approach has two stages: closed-form solution followed by a maximum likelihood criterion–based optimization. Both simulation and real-world experiments have been carried out. The experiment results show that the calibration approach is reliable. This approach will be used in the vehicle detection and tracking system.

Tightly Coupled LIDAR and Computer Vision Integrated System for Vehicle Detection

Computer vision sensors are generally used in current mobile platform–based object detection and tracking systems. However, the application of vision sensors is far from sufficient: clustering, illumination, occlusion, among many other factors, affect the overall performance. In contract, a LIDAR sensor provides range and azimuth measurements from the sensor to the targets. However, the accuracy of its measurements depends on the reflectivity of the targets and the weather. The fusion of camera and active sensors such as LIDAR is being investigated in the context of on-board vehicle detection and tracking.

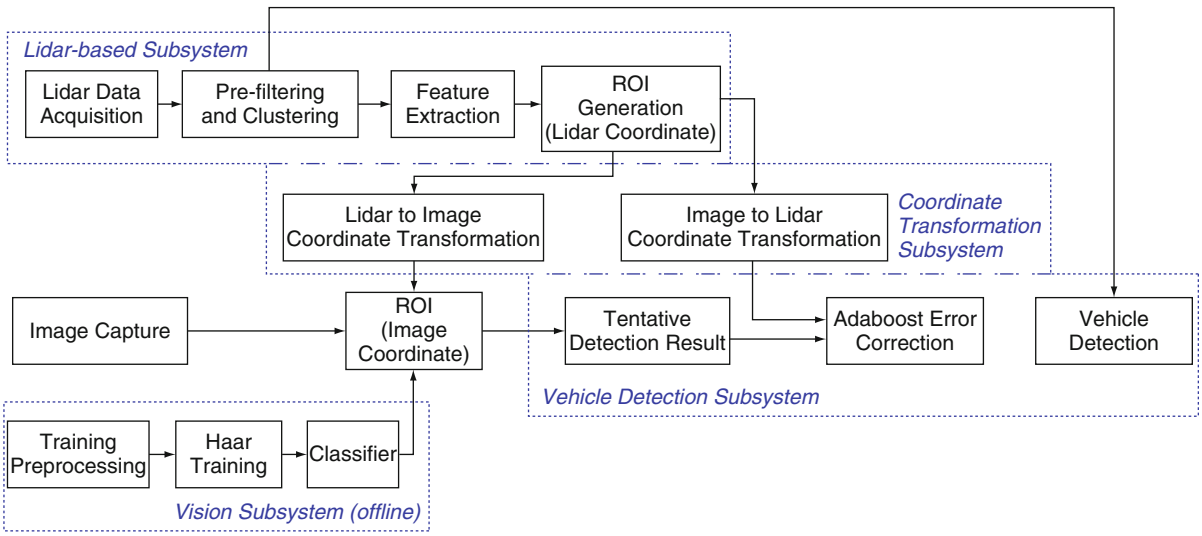
In this section, a tightly coupled LIDAR/CV system is proposed, in which the LIDAR scanning points are used for hypothesizing regions of interest and for providing error correction to the classifier, while the vision image provides object classification information. LIDAR object points are first transformed into image space. ROIs are generated using the LIDAR feature detection method. An Adaboost classifier based on computer vision systems [36] is then used to detect vehicles in the image space. Dimensions and distance information of the detected vehicles are calculated in body-frame coordinates. This approach provides a more complete and accurate map of surrounding vehicles in comparison to the single sensors used separately. One of the key features of this technique is that it uses LIDAR data to correct the Adaboost classification pattern. Moreover, the Adaboost algorithm is utilized both for vehicle detection and for vehicle distance and dimension extraction. Then the classification results provide compensatory information to the LIDAR measurements.

This section is organized as follows: in section "Overview of the Vehicle Detection System" a brief introduction of the vehicle detection system is given. Section "Vision-Based System" describes the vision-based system. The vehicle detection algorithm is introduced in section "Moving Vehicle Detection System". A vehicle tracking approach using particle filter is proposed in section "Vehicle Tracking System". Experimental results of vehicle detection are provided in section "Experiment Results", followed by conclusions and future work in section "Summary and Discussion".

Overview of the Vehicle Detection System

Vehicle detection is the first and fundamental step for any driver assistant system (DAS). With sensors mounted on a moving platform, the detected data change rapidly, making it difficult to extract objects of interest. In the proposed approach, spatial motion of the moving platform has been compensated by using the sensor fusion approach.

Multi-module Architecture The input of vehicle detection system consists of two LIDAR sensors and a single camera, as is shown in Fig. 1. The detection area is covered by two LIDAR sensors overlapping with each other. The camera is placed behind the rearview mirror. The field of view of this camera is fully covered by the LIDAR ranging space.



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 11
Flowchart of the mobile sensing system

Figure 11 presents the flowchart of this vehicle detection system. It consists of four subsystems: a LIDAR-based subsystem, a coordinate transformation subsystem, a vision-based subsystem, and a vehicle detection subsystem.

LIDAR-Based Subsystem *The LIDAR Data Acquisition Module* uses IBEO External Control Unit (ECU) to communicate, collect, and combine data from a pair of LIDAR sensors.

The Prefiltering and Clustering Module aims to transform scan data from distances and azimuths to positions, and cluster the incoming data into segments using a Point-Distance-Based Methodology (PDBM) [12]. If there exists any segment consisting of less than three points, and the distance of this segment is greater than the given threshold, these points are considered as noise. The segment will be disregarded.

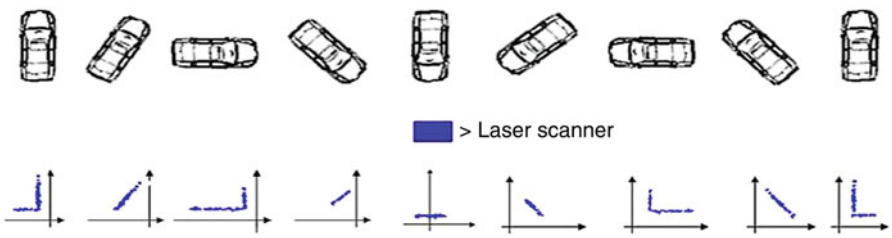
The Feature Extraction Module extracts primary features in the cluster. The main feature information in one segment is its geometrical representation, such as lines and circles. One of the advantages of geometrical features is that they occupy far less space than storing all the scanned points.

A vehicle may have any possible orientation. The contour of a vehicle is constructed by four sides: front, back, left side, and right side. The LIDAR sensor can capture one side, or two neighboring sides, as is shown in Fig. 12.

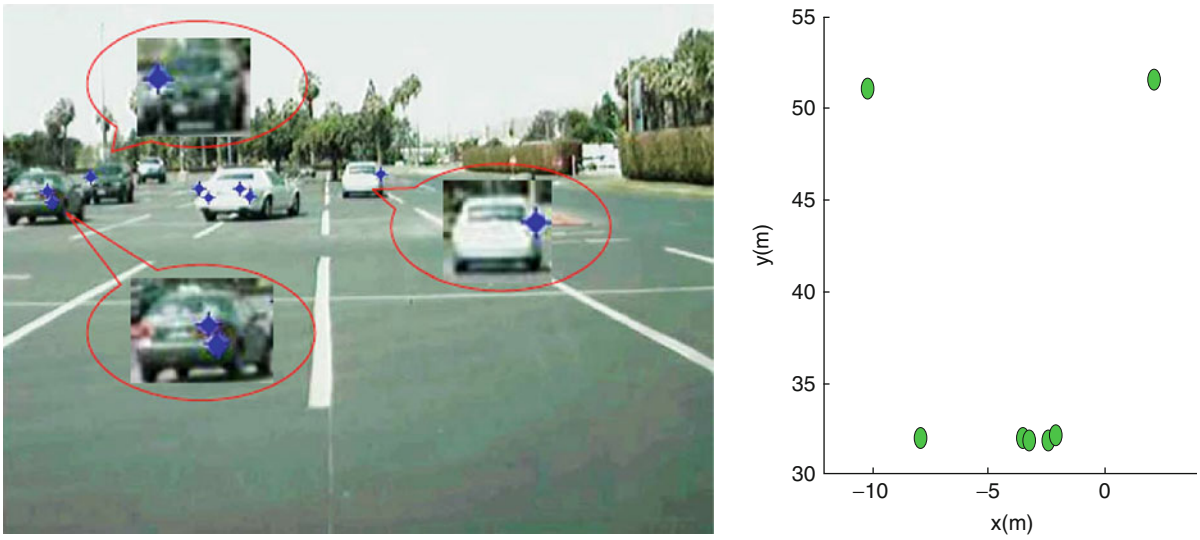
When the object is close to the probe vehicle, the extracted feature provides enough information for object classification. However, if the target is far away, it may be represented by only one or two scanning points. For those objects with only a few LIDAR points, it is difficult to get reliable size, location, and orientation information from the scan data alone. Note that the computer vision image also contains size and orientation information, which can be extracted by the object classification technique. By employing the sensor fusion technology, LIDAR scan data and Adaboost output are complementary to each other.

The ROI Generation Module calculates positions of ROI bounding boxes in LIDAR coordinates. It is worth mentioning that the ROI is not defined by LIDAR points alone, since the scan points of one target may not be able to represent its full dimension. In this algorithm, the width and length of ROI are defined by both LIDAR data points and the maximum dimension of a potential vehicle.

Each ROI is defined as a rectangular area in the image. The bottom of the rectangle is the ground. The top of the rectangle is set to be the maximum height of



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 12
Orientation of the vehicle significantly changes its appearance in the scan data frame. The rectangles show position of the LIDAR



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 13
The LIDAR scan points. (**a**) Points in the LIDAR coordinate system. Each green dot is one scan point. (**b**) These points are transformed to the image frame. Each blue star in the image is one LIDAR point. Vehicles in the red circles are the enlarged image of the detected area

a car. The left and right edges are obtained from the furthest left and right scanning points in a cluster, as well as the typical width of a car.

The LIDAR to Image Coordinate Transformation Module transforms all the LIDAR points into the image frame. The relative position and orientation between LIDAR and camera sensors should be obtained for the transformation. A unique multi-planar LIDAR and computer vision calibration algorithm is described in section "A Novel Multi-Planar LIDAR and Computer Vision Calibration Procedure Using 2D Patterns", which calculates the geometry transformation matrix between "multiple invisible beams" of LIDAR sensors and the camera. The calibration results are used in the sensor fusion system.

During the road test, each LIDAR scan point $\mathbf{P}_l$ is transformed into camera coordinate as $\mathbf{P}_c$. $\mathbf{P}_c$ is then transformed into point $\mathbf{p}_c$ in the image plane. Figure 13 illustrates LIDAR scan points and the transformation results in the image reference system.

After the LIDAR to image transformation coefficients are calculated, ROIs generated in the

LIDAR-based subsystem is converted into image frame. A larger ROI is generated due to inaccuracy of the transformation from LIDAR data to image data.

The Image to LIDAR Coordinate Transformation Module is called to correct Adaboost classification result. More details are given in the following sections.

Vision-Based System

Object classification from the hypothesized ROIs is required for vehicle detection purpose. Feature representations are used for object classifiers by an Adaboost algorithm [36]. Viola et al. proposed that the object is detected based on a boosted cascade of feature classifiers, which performs feature extraction and combines features such as simple weak classifiers to a strong one.

The Adaboost classifier requires off-line training using target as well as nontarget images. In the vehicle detection applications, the target images, i.e., the rearview of vehicles, are called positive samples; while the non-vehicles are named as negative samples. Figure 14 illustrates some of the positive as well as negative samples in the training dataset. Training samples are taken from both Caltech vehicle image dataset [37] and video collected by the test vehicle. The positive samples include passenger cars, vans, trucks, and trailers. The negative sample sets include roads, traffic signs, buildings, plants, and pedestrians.

Image Training Preprocessing Both the positive and negative samples are used by computer vision system for data training. All the samples are originally colored images. In order to remove the effects of various illumination conditions and camera differences, gray-level transformation is required as a preprocessing step. The gray-level normalization method is applied to the whole image dataset, which transforms gray level of the image to be in [0, 1] domain. The color image is transformed by [38]:

$$I^*(x,y) = \frac{I(x,y) - I_{\min}}{I_{\max} - I_{\min}} \tag{11}$$

where $I(x,y)$ is the intensity of pixel (x,y), $I_{\min}$ and $I_{\max}$ are the minimum and maximum values in this image, respectively. $I^*(x,y)$ is the normalized gray-level value.

The next step is to normalize the sizes of all the positive samples. It is implemented before training

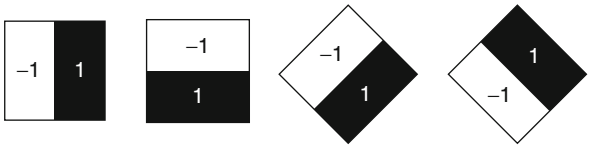
Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 14
Positive and negative samples used in Adaboost training

since different resolutions may cause different number of features to be counted. The sizes of the normalized positive samples determine the minimum size of objects that can be detected [39]. In this test, the normalized image size is set as 25 × 25 pixels.

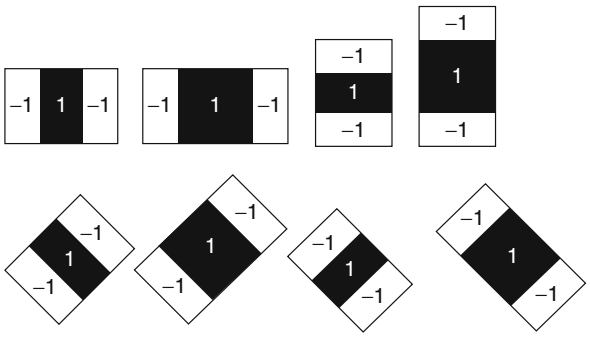
Haar Training In the vision-based system, Haar-like features are used to classify objects [36]. This approach combines more complex classifiers in a "cascade" which quickly discard the background regions while spending more computation on the Haar-like area [36].

More specifically, 14 feature prototypes are utilized for the Haar training [40]. These features represent characteristic properties like edge, line, and symmetry. The features prototypes can be grouped into three categories:

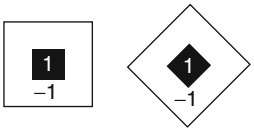
- Edge features: The difference between sums of the pixels within two rectangular regions



- Line features: The sum within two outside rectangular regions subtracted from the sum in the center rectangular



- Center-surround features: The difference between the sums of the center rectangular and the outside area



Here black areas with "–1" have negative weights, and white areas with "1" have positive weights.

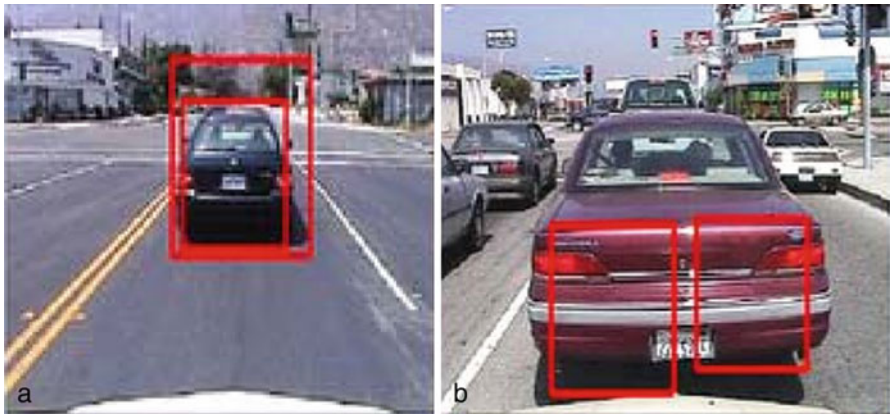
After the weak classifier has been trained at each stage, the classifier is able to detect almost all the targets of interest while rejecting certain nontarget objects. A cascade of classifiers is generated to form a decision tree. The training process consists of totally 15 stages. Each stage is trained to eliminate 60% of the nonvehicle patterns, and the hit rate (HR) in each stage is set to be 0.998. Therefore, the total false alarm rate (FAR) for this cascade classifier is supposed to be $0.4^{15} \approx 1.07e-06$, and the hit rate should be around $0.998^{15} \approx 0.97$.

Moving Vehicle Detection System

The Adaboost algorithm designs a strong classifier that can detect multiple objects in the given image. However, there is no guarantee that this strong classifier is optimal for the object detection. In contrast to the classic Adaboost algorithm, in this test there is only one vehicle in each ROI defined by the LIDAR clustering algorithm. Therefore, it is not necessary for the Adaboost algorithm to detect several possible targets in one ROI. A classification correction technique is proposed to utilize the LIDAR scanning data to reduce redundancy in the Adaboost detection results.

There are two kinds of redundancy errors in the classification results. Figure 15 gives some examples of these two cases. Both have detected more than one object, while the ground truth is that there is only one vehicle. One kind of error is that the Adaboost detects two possible targets, while the area of the smaller one is almost covered by the larger one, as is shown in Fig. 15a. Another error shown in Figure 15b is that all the detected areas belong to the same object, while none of them cover the full body of the target.

Suppose in the i-th ROI, there exists a LIDAR point cluster $\Re_i$, which has the following features in LIDAR coordinate system: c_i^{LIDAR} as the center of the cluster, w_i^{LIDAR} as the width of the object, and l_i^{LIDAR} as the possible length of the vehicle. On the image side, there are detected target candidates denoted as d_1, d_2,. . ., d_n. Initially, d_1, d_2,. . ., d_n are transformed from image coordinate to camera coordinate, then to LIDAR coordinates.



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 15
Two kinds of redundancy errors in Adaboost detection. (**a**) Two bounding boxes overlap on the same target. (**b**) Two bounding boxes detected with no overlap on the same target

The scan points in LIDAR coordinates are denoted as D_1, D_2,..., D_n. The j-th candidate D_j has center $c_{i,j}^{DETECT}$, width $w_{i,j}^{DETECT}$, and height $h_{i,j}^{DETECT}$. In LIDAR coordinate frame, two vectors are defined as $\mathbf{m}_i^{LIDAR} = \left(c_i^{LIDAR}, w_i^{LIDAR}\right)$ and $\mathbf{m}_i^{DETECT} = \left(c_{i,1}^{DETECT}, w_{i,1}^{DETECT}, \cdots, c_{i,n}^{DETECT}, w_{i,n}^{DETECT}\right)$. Here $\mathbf{m}_i^{LIDAR}$ is the size information obtained by LIDAR, and $\mathbf{m}_i^{DETECT}$ consists of measurements in LIDAR coordinate systems which are transformed from the image reference system. The coefficient $\mathbf{w}_i$ for mapping multiple target areas to the LIDAR information satisfies:

$$\mathbf{w}_i = \arg\min \left\| \mathbf{m}_i^{LIDAR} - \sum\nolimits_{j=1:n} \mathbf{w}_{i,j} \mathbf{m}_{i,j}^{DETECT} \right\| \tag{12}$$

where $\|.\|$ is the Euclidean norm. $\mathbf{w}_i$ is used as a weight to recalculate the detected area, which is a combination of all the detected objects in one ROI.

LIDAR scanning points and Adaboost classification results are then combined to generate a complete map of vehicles. A summary of the vehicle detection process is given here:

Algorithm 1: Vehicle Detection. Given LIDAR points cluster $\Re_i$ with features $\left(c_i^{LIDAR}, w_i^{LIDAR}, l_i^{LIDAR}\right)$, detected target candidate d_1, ..., d_n in the image;

if no object detected in the ROI
  enlarge ROI and search again
else
  define $\mathbf{m}_i^{LIDAR} = \left(c_i^{LIDAR}, w_i^{LIDAR}\right)$
  transform d_1, ..., d_n in the image coordinate frame to LIDAR reference frame
  define $\mathbf{m}_i^{DETECT} = \left(c_{i,1}^{DETECT}, w_{i,1}^{DETECT}, \cdots, c_{i,n}^{DETECT}, w_{i,n}^{DETECT}\right)$
  calculate the weight vector which minimize $\left\| \mathbf{m}_i^{LIDAR} - \sum \mathbf{w}_{i,j} \mathbf{m}_{i,j}^{DETECT} \right\|$
end if
The detected vehicle is located at $\sum_{j=1:n} \mathbf{w}_{i,j} c_{i,j}^{DETECT}$.

The vehicle detection system proposed in this section can be summarized as:

- The Adaboost classifier training is implemented off-line with both positive and negative samples.
- ROI is defined by the LIDAR scan data. No more than one vehicle is assumed to exist in each ROI.
- Use the Adaboost classifier to make a preliminary vehicle detection.
- LIDAR data is used to correct the Adaboost redundancy error, and to merge detected areas in one ROI.
- Combine the Adaboost detected area (in LIDAR coordinate) and the LIDAR output to generate a complete vehicle distance and dimension map.

Vehicle Tracking System

LIDAR and computer vision sensors are integrated in a probabilistic manner for vehicle tracking. A sampling

importance resampling (SIR) particle filter is used as the tracker, which is a sophisticated model estimation technique [41]. Unlike the commonly used Kalman filter and extended Kalman filter (EKF), this particle filter does not assume that the linear dynamic system is perturbed by Gaussian noise. The key idea of particle filter is to represent the estimation by a set of random samples (they are called *particles*) with associated weights.

Particle Filter Let $\mathbf{x}_t^{(i)}$ be the i-th sample of the position and velocity of the target at time t, $i = 1, 2, \cdots, N_s$, where N_s is the total number of samples or particles in the particle filter. $w_t^{(i)}$ is the i-th weight at time $_t$ associated with $\mathbf{x}_t^{(i)}$. The procedure of SIR particle filter is defined as follows:

1. Initial Particle Generation
   Generate N_s particles $\left\{\mathbf{x}_0^{(i)},\ w_0^{(i)}\right\}$, $i = 1, 2, \cdots, N_s$. Here $\mathbf{x}_0^{(i)}$ is obtained from vehicle detection results and $w_0^{(i)} = 1/N_s$.
2. Particle Updating
   For each particle $\mathbf{x}_{t-1}^{(i)}$ at time $t-1$, generate a particle $\mathbf{x}_t^{(i)}$ at time t. This step corresponds to the prediction step in Kalman filter and EKF. However, in Kalman filter and EKF, at time t the state is updated only once. Here in the particle filter, each of the particles should be updated, so totally N_s particle updating calculations are implemented. In this system, $\mathbf{x}_t^{(i)} = \begin{bmatrix}\mathbf{p}_t^{(i)} & \mathbf{v}_t^{(i)}\end{bmatrix}$ is the sample of the position and velocity, so a linear model is used to update the particles:
   $$\mathbf{p}_t^{(i)} = \mathbf{p}_{t-1}^{(i)} + \mathbf{v}_{t-1}^{(i)} T + n \tag{13}$$
   where T is the time interval and n is the noise.
3. Particle Weighting
   Each particle $\mathbf{x}_0^{(i)}$ at time t is associated with the weight $w_t^{(i)}$, which is also called the *importance factor*. The weight at time t is a function of the weight at time $t-1$, and the probability function of measurements and states at time t. The importance factor is commonly calculated as [41]:
   $$w_t^{(i)} = w_{t-1}^{(i)} p\left(\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \tag{14}$$
   where $\mathbf{z}_t$ is the measurement at time t. In the proposed sensor fusion system, both LIDAR and computer vision sensors are utilized for vehicle tracking. Therefore, the measurement is $\mathbf{z}_t = \left\{ {}^l\mathbf{z}_t;\ {}^c\mathbf{z}_t \right\}$, where ${}^l\mathbf{z}_t$ is the output of LIDAR and ${}^c\mathbf{z}_t$ is the measurement of the camera. This step corresponds to the update step in Kalman filter and EKF. The probability $p\left(\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right)$ will be discussed in the following subsection.
4. Resampling
   A common problem with the particle filter is the degeneracy, which is the phenomenon that after a few iterations only one particle has non-negligible weight [41]. $\widehat{N_{eff}}$ has been defined to measure the degeneracy, which is calculated as [41]:
   $$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N_s} \left(w_t^{(i)'}\right)^2} \tag{15}$$
   where $w_t^{(i)'}$ is the normalized weight of the i-th particle at time t with $w_t^{(i)'} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_s} w_t^{(i)}}$. If $\widehat{N_{eff}}$ is less than a given threshold N_T, the degeneracy is detected.

   Resampling is performed at each iteration. It is designed to eliminate particles that have small weights and replicate particles that have large weights. Particles that have large weights are considered to be the "good" particles while the particles with small weights are "bad" particles. The resampled weights are set as $w_t^{(i)} = 1/N_s$.

   After obtaining $w_t^{(i)}$, the posterior filtered density can be approximated as [41]:
   $$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} w_k^i \delta\left(\mathbf{x}_t - \mathbf{x}_t^{(i)}\right) \tag{16}$$

The Sensor Model The sensor model describes the process by which sensor measurements are made in the physical world. It relates sensor output to the state of the vehicle. In the vehicle detection application it is defined as the conditional probability $p\left({}^l\mathbf{z}_t;\ {}^c\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right)$, which is the probability of LIDAR and computer vision sensor measurements given the state of the vehicle.

A LIDAR sensor model is described in [42], which represents the probability as a mixture of four distributions corresponding to four types of measurement errors: the small measurement noise, errors due to unexpected objects or obstacles, errors due to failure to detect objects, and random unexplained noise.

Let $z_t^{(k)*}$ denote the true distance to an obstacle, $z_t^{(k)}$ denote the recorded measurement, and $z_{\max}$ denote the maximum possible reading. The small measurement error is defined as a Gaussian distribution p_{hit} over the range $[0, \quad z_{\max}]$ with mean $z_t^{(k)*}$ and standard deviation σ_{hit}.

The LIDAR detection zone is often blocked by the moving vehicles, which leads to the measurement whose length is shorter than the true length. This particular type of measurement error is modeled by a truncated exponential distribution p_{short} with the coefficient λ_{short}.

LIDAR sometimes fails to detect obstacles due to low reflectivity of the target. The errors due to failure to detect objects is defined as a pseudo point-mass distribution $p_{\max}$ centered at $z_{\max}$.

Finally, unexplainable measurements may be returned by the LIDAR sensor, which is caused by interference. This type of error is modeled by a uniform distribution p_{rand} over the entire measurement range.

$p\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right)$ is calculated as a combination of the four types of errors as [42]:

$$\begin{aligned} p\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = {} & \alpha_{\text{hit}} p_{\text{hit}}\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ & + \alpha_{\text{short}} p_{\text{short}}\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ & + \alpha_{\max} p_{\max}\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ & + \alpha_{\text{rand}} p_{\text{rand}}\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \end{aligned} \tag{17}$$

where α_{hit}, α_{short}, $\alpha_{\max}$, and α_{rand} are the weights for p_{hit}, p_{short}, $p_{\max}$, and p_{rand}, respectively. $\alpha_{\text{hit}} + \alpha_{\text{short}} + \alpha_{\max} + \alpha_{\text{rand}} = 1$. The parameters in Eq. 17 are commonly used as the a priori information, which are obtained by data training.

A camera weight model is proposed in [43] as:

$$p\left({}^c\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = \begin{cases} S & \text{the object is in the camera detection zone.} \\ 1-S & \text{the object is out of the camera detection zone.} \end{cases} \tag{18}$$

where S is a constant, $0 \le S \le 1$. This model is based on the assumption that the camera is able to detect all the objects in the detection zone.

The sensor fusion probability model is calculated based on the LIDAR probability model as well as the camera probability model. It is proposed in [43] that ${}^l\mathbf{z}_t$ and ${}^c\mathbf{z}_t$ are independent measurements. So

$$p\left({}^l\mathbf{z}_t; {}^c\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = p\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) p\left({}^c\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \tag{19}$$

However, in the proposed sensor fusion system, LIDAR and the camera are not two independent sensors. They have been calibrated to observe the same target, and the geometric relationships are given as a priori information. Moreover, the classification result of the camera is corrected by the LIDAR output.

In this section, a novel sensor fusion probability model is proposed. As defined in [42], LIDAR tracking process is modeled as a mixture of four types of errors: the small measurement noise, unexpected objects detection error, detection failure error, and random unexplained noise. In the field test, small measurement noise error is found to be the exclusive error source of the LIDAR sensor. The other types of errors are removed by integration of LIDAR and camera. The LIDAR tracking model is [42]:

$$p\left({}^l\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = \begin{cases} \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{\left({}^l\mathbf{z}_t - {}^l\mathbf{z}_t^*\right)^2}{2\delta^2}\right) & 0 \le {}^l\mathbf{z}_t \le z_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

The computer vision-based vehicle tracking is implemented by KLT tracking [36, 47]. A function $s\left(x_{t(i)}^m\right)$ is used to determine if a predicted corner $x_{t(i)}^m$ is close to an observed corner $z_{t(i)}$. $s\left(x_{t(i)}^m\right)$ is defined as $s\left(x_{t(i)}^m\right) = -\sum_{j=1}^{M} \exp\left(d_{t(i,j)}^m\right)^2$, where M is the total number of corners, $\left(d_{t(i,j)}^m\right)^2 = \left\| z_{t(j)} - x_{t(i)}^m \right\|^2$ is the Euclidean distance between i-th detected corner of the m-th particle $x_{t(i)}^m$ and j-th extracted corner $z_{t(j)}$ [47]. The camera model is defined as [47]:

$$p\left({}^c\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = \exp\left(-\sum \left(s\left(x_{t(i)}^m\right) - 1\right)^2\right) \tag{21}$$

Finally, the sensor fusion tracing system has totally three measurement situations: (1) the vehicle is tracked by both the LIDAR and the camera. In this case, the single sensor tracking error is eliminated by sensor integration technique proposed in section "Moving

Vehicle Detection System"; (2) the vehicle is out of camera detection zone, so it is tracked by the LIDAR alone; and (3) the vehicle is tracked by the camera but not detected by the LIDAR sensor due to factors such as distance or weak reflection. The sensor fusion model is given as:

$$p\left({}^{l}\mathbf{z}_t;{}^{c}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) = \begin{cases} \alpha p\left({}^{l}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) + \beta p\left({}^{c}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ \text{target is tracked by both LIDAR and camara} \\ p\left({}^{l}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ \text{target is tracked by LIDAR alone} \\ p\left({}^{l}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right) \\ \text{target is tracked by camara alone} \end{cases} \tag{22}$$

where the weight $0 \le \alpha \le 1$ and $0 \le \beta \le 1$ are two coefficients obtained by data training, $\alpha + \beta = 1$, which allows to balance the LIDAR and computer vision information.

The weight $w_t^{(i)}$ can be calculated using Eq. 14. The particle filter is summarized in Algorithm 2. Unlike the Kalman filter or EKF, particle filter can track vehicle state with multi-model or arbitrary distributions.

Algorithm 2: Particle Filter for Sensor Fusion Systems. **Input:** $\left\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$, $i = 1, 2, \cdots, N_s$: set of weighted particles at time $t-1$

$\mathbf{z}_t = \left({}^{l}\mathbf{z}_t; {}^{c}\mathbf{z}_t\right)$: LIDAR and computer vision measurement at time t

Output: $\left\{x_t^{(i)}, w_t^{(i)}\right\}$, $i = 1, 2, \cdots, N_s$: set of weighted particles at time t

Process:

for $i = 1$ to N_s do

Predict $\mathbf{x}_t^{(i)}$ as $\mathbf{p}_t^{(i)} = \mathbf{p}_{t-1}^{(i)} + \mathbf{v}_{t-1}^{(i)} T + n$

$w_t^{(i)} = w_{t-1}^{(i)} p\left(\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right)$

end for

calculate $\widehat{N_{eff}}$

if $\widehat{N_{eff}} < N_T$

for $i = 1$ to N_s do

computer $w_t^{(i)}$ using Eq. 14, in which $p\left({}^{l}\mathbf{z}_t; {}^{c}\mathbf{z}_t \middle| \mathbf{x}_t^{(i)}\right)$ is given in Eq. 20

update the particle with $\left\{x_t^{(i)}, 1/N_s\right\}$

end for

else

$Z = \sum_{i=1}^{N_s} w_t^{(i)}$

for $i = 1$ to N_s do

update the particle with $\left\{x_t^{(i)}, Z^{-1} w_t^{(i)}\right\}$

end for

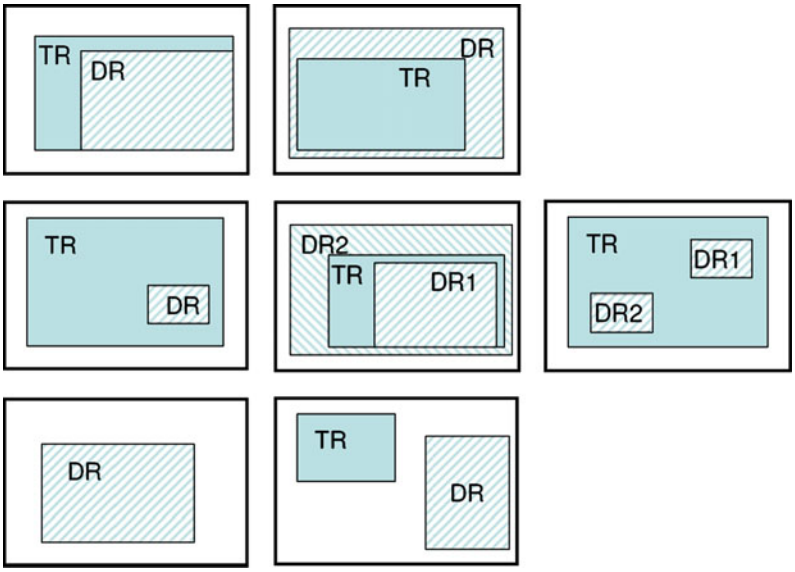
end if

Experiment Results

The experiment results of vehicle detection are discussed in this section. To evaluate the performance of this system, a dataset of 377 images was used for training and testing. These images are taken from both the Caltech vehicle image dataset [37] and the video samples collected by probe vehicles. Another test dataset consists of 526 images with synchronized scanning data that are used for performance evaluation. The test dataset was recorded in a local parking lot on different days during different seasons.

Hit rate (HR), false alarm rate (FAR), and region detection rate (RDR) are used to evaluate the performance of this system. Here HR is the number of detected vehicles over total number of vehicles. RDA denotes the percentage of "real" vehicle detection rate. A "real" vehicle detection is that majority area of the vehicle is covered by a rectangle, and there is only one rectangle that covers this object. Therefore, a target that is hit may not be a region "really" detected; and a region detected object is always a hit. The higher the RDR is, the more accurate the detection result will be. Figure 16 illustrates several cases for hit, false alarm, and region detection. In this figure, TR represents the target region in the image, and DR is the detected region by the classifier.

Table 2 gives the detection performance of Adaboost classifier (detection with camera only), the classic LIDAR–camera sensor fusion system, and the proposed LIDAR and computer vision-based detection and error correction approach. Here in both the classic sensor fusion technique and the proposed approach, LIDAR data are utilized for ROI generation. The difference lies in the fact that in the proposed approach LIDAR data help correct the classification result. Table 2 shows that this approach both improves the hit rate and reduces the false alarm rate in comparison with Adaboost classifier. Compared with the classic LIDAR–camera fusion system, this approach improves the



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 16
Three target detection cases. The first row is region detected, the second row is hit but not region detected, and the third row is false alarm

Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Table 2 Detection result

Type	HR (%)	FAR (%)	RDR (%)
Adaboost	84.17	3.27	78.00
Classic LIDAR–camera fusion	91.33	1.78	84.85
Proposed approach	91.33	1.78	89.32

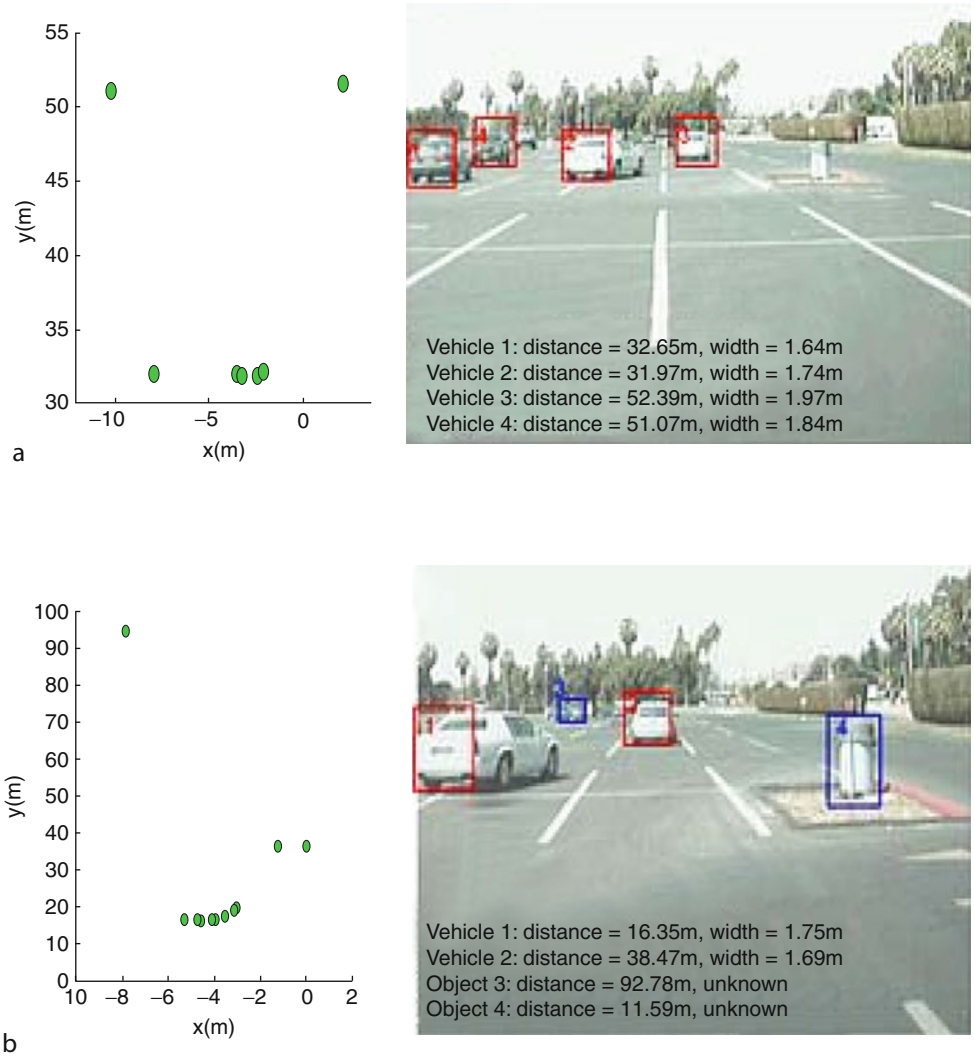
region detect rate from 84.85% to 89.32%, since for each hit but not accurately covered object the LIDAR scanning data helps to recomputed position of the target. Most of the overlapping or partial target detection areas are merged during the LIDAR correction process.

Figure 17 illustrates some of the vehicle detection results. The left column presents LIDAR scan points, and the right column illustrates camera images with information from the sensor fusion system. In Figure 7a, all the vehicles are detected and are marked with a rectangle. Figure 7b shows that the classifier found two vehicles only, which are bounded in a red rectangle. The other two ROIs (shown in blue rectangles) are classified to have non-vehicle objects. In fact, one of them is a trash can. The other is a vehicle at a distance of 92.78 m. This vehicle is too far away and too small in the image for the classifier to recognize.

During the test, the hit rate decreases when the distance between the probe vehicle and the target vehicles increases. The targets are detected frame by frame. Therefore, the target vehicles may not be recognized by the classifier in certain frames even if it was recognized in the last frame. Vehicle tracking technique helps solve this problem. By running a particle filter–based tracking algorithm, the target is initially detected in the initial frame or in several initial frames, after which it is tracked in the following frames. This approach both improves the detection accuracy and reduces the required amount of calculation. HR and RDR will be further improved by vehicle tracking.

Summary and Discussion

A novel vehicle detection system has been proposed based on tightly integrating LIDAR and computer vision sensors. Distances to the objects are first defined by the LIDAR sensors, and then the object is classified based on computer vision images. In addition, data from these two complementary sensors are combined for classifier correction and vehicle detection.



Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for. Figure 17
LIDAR scan points and the final vehicle detection results

The experimental results have indicated that, when compared with image-based and classic sensor fusion–based vehicle detection systems, this approach has a higher hit rate and a lower false alarm rate. It is quite useful for modeling and prediction of the traffic conditions over a variety of roadways. This system may be used in future autonomous navigation systems.

Conclusions and Future Work

This entry presents a multi-sensor equipped vehicle detection system that was developed to specifically obtain the state of surrounding vehicles. It involves the development of a tightly coupled LIDAR and computer vision system, calibration of a pair of multi-planar LIDAR sensors and the camera system, and the methodology of sensor fusion-based vehicle detection technique.

This section provides a brief summary of this entry, as well as the possible future work.

Summary

Automatic vehicle detection techniques are becoming an essential part of our daily lives. They open up many potential opportunities but they also come with

challenges in terms of sensing capability and accuracy. In this entry, the problem of vehicle detection is addressed, and some novel approaches have been demonstrated to solve the problem in a traffic environment.

The goal of this research is to provide a solution to measure the state of surrounding vehicles. State of the vehicle includes position, orientation, speed, and acceleration. Sensor fusion techniques are utilized to provide a direct measurement of the state. A variety of sensors have been used in this entry, including LIDAR and computer vision. The goal is to quantitatively show that the integration of sensors provides a more accurate and effective estimation of the vehicle state. The proposed system has successfully met this goal.

The developed multi-planar LIDAR and computer vision sensor calibration approach, as to the author's best knowledge, is the first calibration method for an "invisible-beam" multi-planar LIDAR and a camera. In comparison to the commonly used calibration methods that require a infrared camera to "see" the LIDAR beams or a special designed calibration shape, this approach is easy to implement with low cost. It has been theoretically and experimentally proven to be able to estimate the geometric relationships between the two sensors.

Based on this unique calibration method, a sensor fusion-based vehicle detection system is designed and implemented. It consists of three major components: (1) ROIs are generated by the LIDAR sensor; (2) vehicle classification using a computer vision-based Adaboost algorithm, and (3) vehicle position is verified using the output of the LIDAR sensor. A vehicle tracking model is also presented in this entry, which uses a joint probability model-based particle filter to predict state of the vehicle. The experiment result shows that the designed sensor fusion system achieves higher detection rate and lower positive as well as negative error rates, when compared with a single sensor-based detection method. The positions of detected vehicles have been represented in vehicle coordinates to generate a local traffic map.

Taken together, the tests in this entry demonstrate that a good vehicle detection performance can be achieved using a LIDAR and computer vision sensor-based moving platform. Such results are especially important for vehicle navigation systems, as well as traffic surveillance systems that are equipped with multiple sensors.

Future Directions

Although a sensor fusion system is developed in this entry for the on-board vehicle detection application, it is believed that the introduction of sensor fusion-based system in the automobile industry is still couple of years away. In the future driver assistance systems, sensor fusion techniques can be employed to support, even replace the driver. Moreover, falling costs of sensors, such as RADAR, GPS, inertial sensors (INS), and LIDAR, combined with increasing image processing capability provides a bright future for on-board intelligent transportation applications.

Disclaimer

Much of the material for this entry comes from the 2010 dissertation of Lili Huang, at the University of California-Riverside (see [44]). Portions of this entry have also appeared in [45] and [46].

Bibliography

1. Proper AT, Cheslow MD (1997) ITS benefits: continuing successes and operational test results. FHWA-JPO-98-002, Oct 1997
2. Mimbela L, Klein LA (2000) Summary of vehicle detection and surveillance technologies used in intelligent transportation systems. Report to Federal Highway Administrations (FHWA) Intelligent Transportation Systems Joint Program Office
3. Apogee/Hagler Bailly (1998) Intelligent transportation systems: real world benefits. FHWA-JPO-98-018, Jan 1998
4. Kim MY, Cho H, Lee H (2004) An active trinocular vision system for sensing mobile robot navigation environments. In: Proceedings of the IEEE intelligent robots and systems, Sendai, pp 1698–1703
5. Jung YK, Ho Y-S (1999) Traffic parameter extraction using video-based vehicle tracking. In: IEEE/IEEJ/JSAI international conference on intelligent transportation systems, Tokyo, pp 764–769
6. Baltzakis H, Argyros A, Trahanias P (2003) Fusion of laser and visual data for robot motion planning and collision avoidance. Mach Vision Appl 12:431–441
7. Liu Y, Emery R (2001) Using EM to learn 3D environment models with mobile robots. In: Proceedings of the 18th international conference on machine learning, Williamston, 2001
8. Jokinen O (1999) Self-calibration of a light striping system by matching multiple 3-D profile maps. In: Proceedings of the second international conference on 3-D digital imaging and modeling, Ottawa, pp 180–190
9. Hayashibe M, Nakamura Y (2001) Laser-pointing endoscope system for Intra-operative 3D geometric registration. In: Proceedings of the international conference on robotics and automation, Seoul, pp 1543–1548

10. Mahlisch M, Schweiger R, Ritter W, Dietmayer K (2006) Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection. In: Proceedings of intelligent vehicles symposium, Meguro-Kupp, pp 424–429
11. Unnikrishnan R, Hebert M (2005) Fast extrinsic calibration of a laser rangefinder to a camera. Technical report CMU-RI-TR-05-09, Robotics Institute, Carnegie Mellon University, July 2005
12. Zhang Q, Pless R (2004) Extrinsic calibration of a camera and laser range finder (Improves camera calibration). In: IEEE proceedings of international conference on intelligent robots and systems, Sendai, pp 2301–2306
13. DARPA urban challenge http://www.darpa.mil/grandchallenge/index.asp
14. Mahmassani HS, Haas C, Zhou S, Peterman J (1998) Evaluation of incident detection methodologies. Technical Report FHWA/TX-00/1795-1, Center of Transportation Research, The University of Texas at Austin, Oct 1998
15. University of Central Florida (2007) I2Lab, TeamUCF – DARPA urban challenge. Technical Report, June 2007
16. Thrun S (2003) Learning occupancy grid maps with forward sensor models. Autonomous Robots 15(2):111–127
17. SICK USA, see http://www.sick.com/us/en-us/home/Pages/Homepage1.aspx
18. HOKUYO UXM-30LN LIDAR, see http://www.hokuyo-aut.jp/02sensor/07scanner/uxm_30ln.html
19. The laser scanner product overview, see http://www.ibeoas.com/english/products.asp
20. Velodyne HDL-64E LIDAR, http://www.hizook.com/blog/2009/01/04/velodyne-hdl-64e-laser-rangefinder-lidar-pseudo-disassembled
21. Whittaker WR (2005) Red team DARPA grand challenge 2005 technical paper. Carnegie Mellon University, CSC Technical Report, Aug 2005
22. Cobzas D, Zhang H, Jagersand M (2002) A comparative analysis of geometric and image-based volumetric and intensity data registration algorithms. In: IEEE international conference on robotics and automation (ICRA), Washington, DC, 2002
23. Wasielewski S, Strauss O (1995) Calibration of a multi-sensor system laser rangefinder/camera. In: Proceedings of intelligent vehicles symposium, Detroit, pp 472–477
24. Neira J, Tard JD, Horn J, Schmidt G (1999) Fusing range and intensity images for mobile robot localization. IEEE Trans Robot Autom 15(1):76–84
25. Toyh D, Aach T (2003) Detection and recognition of moving objects using statistical motion detection and Fourier descriptors. In: 12th international conference on image analysis and processing, Mantova, pp 430–435
26. Oren M, Papageorgiou C, Sinha P (1997) Pedestrian detection using wavelet templates. In: IEEE proceedings on computer vision and pattern recognition, San Juan, 1997
27. Premebida C, Monteiro G, Nunes U, Peixoto P (2007) A lidar and vision-based approach for pedestrian and vehicle detection and tracking. In: IEEE intelligent transportation systems conference, Seattle, pp 1044–1049
28. Fardi B, Schuenert U, Wanielik G (2005) Shape and motion-based Pedestrian detection in infrared images: a multi sensor approach. In: Proceedings of the IEEE intelligent vehicles symposium Las Vegas, pp 18–23
29. Milch S, Behrens M (2001) Pedestrian detection with radar and computer vision. In: Proceedings of the conference on progress in automobile lighting, Darmstadt, 2001
30. Zhang Z (2000) A flexible new technique for camera calibration. IEEE Trans Patt Anal Mach Intell 22:1330–1334
31. Bouguet J-Y (2003) Camera calibration toolbox for matlab, 2003
32. Faugeras OD, Toscani G (1987) Camera calibration for 3D computer vision. In: Proceedings of the international workshop on industrial applications of machine vision and machine intelligence, Silken, Japan, pp 240–247
33. Ricolfe Viala C, Sanchez Salmeron AJ (2004) Performance evaluation of linear camera calibration techniques. In: IEEE Proceedings on world automation congress, Spain, vol 18, pp 49–54
34. Levenberg K (1944) A method for the solution of certain problems in least squares. Quart Appl Math 2:164–168
35. More J (1977) The Levenberg-Marquardt algorithm, implementation and theory. In: Numerical analysis. Lecture notes in mathematics, vol 630. Springer, Berlin
36. Viola P, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE computer vision and pattern recognition, Hawaii, pp 511–518
37. Caltech vehicle image dataset, see http://www.vision.caltech.edu/htmlfiles
38. Haselhoff A, Kummert A, Schneider G (2007) Radar-vision fusion with an application to car-following using an improved adaboost detection algorithm. In: IEEE intelligent transportation systems conference, Seattle, pp 854–858
39. Tutorial OpenCV Haartraining, see http://sourceforge.net/projects/opencvlibrary
40. Lienhart R, Kuranov A, Pisarevsky V (2003) Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: 25th pattern recognition symposium, Madgeburg, Sep 2003, pp 297–304
41. Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50(2):174–188
42. Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. The MIT Press, Cambridge, MA
43. Kurazume R, Yamada H, Murakami K, Iwashita Y, Hasegawa T (2008) Target tracking using SIR and MCMC particle filters by multiple cameras and laser range finders. In: IEEE/RSJ international conference on intelligent robots and systems, Nice, pp 3838–3844
44. Huang L (2010) LIDAR, Camera and inertial sensors based navigation techniques for advanced intelligent transportation systems, PhD dissertation, University of California Riverside
45. Huang L, Barth M (2009) A novel multi-planar LIDAR AND computer vision calibration procedure using 2D patterns for automated navigation. In: IEEE intelligent vehicles symposium, Xi'an, pp 117–122

46. Huang L, Barth M (2009) Tightly-coupled LIDAR and computer vision integration for vehicle detection. In: IEEE intelligent vehicles symposium, Xi'an, pp 604–609
47. Dore A, Beoldo A, Regazzoni CS (2009) Multitarget tracking with a corner-based particle filter. In: IEEE 12th international conference on computer vision workshops, Kyoto, pp 1251–1258

Vehicle Dynamics and Performance

Yimin Gao
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

Article Outline

Glossary

Braking forces The forces acting on the contact area of the running wheels and ground, generated by the brake system.

Braking performance The vehicle behavior underlying braking, braking distance, and direction stability.

Braking system A vehicle subsystem that is used to slow the vehicle quickly.

Fuel consumption Fuel consumed in unit traveling distance.

Power plants The machines that supplies power for propelling vehicle.

Tractive effort The thrust force acting on the contact area of running wheels and ground that push the vehicle forward.

Transmissions Mechanical devices that transmit the powers of the power plants to vehicle wheels.

Vehicle performance The capability of a vehicle, in terms of speed, acceleration, and gradeability.

Vehicle resistance The forces that is against the vehicle motion.

Definition of the Subject

Vehicle system is a complex system that includes many mechanical and electric components and operates in very different traffic environments. There are many requirements, including traffic conditions, mission requirements, energy supplies, environmental protection, cost, etc. Vehicle dynamic and performance analysis supplies the basic methodologies for vehicle performance evaluation, design principle, and quantitative computation methods for system and component design.

Introduction

Vehicle dynamics and performance are broad topics that deal with vehicle's drivability, fuel economy, braking performance, handling characteristics, noise, vibration and harshness (NVH), etc. The research for improving vehicle dynamics and performance never ceased in 100 years since vehicles have been invented. The research in this area has been developed for the purposes of basic understanding of the system operation behaviors, system and components design, and development of more advanced control technologies, such as engine control, transmission control, traction control, braking control, vehicle stability control, etc. In recent years, more efficient and clean vehicle technologies have been developed quickly, especially, electric propulsion, fuel cell, and hybrid technologies. These require the research on vehicle dynamics and performance being extended beyond the scope of conventional vehicles that has focused on gasoline and diesel power vehicles. However, although there are many differences between conventional and electric-based vehicles, they share some similarities of drivability and braking performance. The fundamentals of vehicle dynamics and performance established for conventional vehicles are still, to a great degree, valid. This section reviews the fundamentals of vehicle's dynamics and performance for providing reference for electric, hybrid electric, and fuel cell vehicle design. The following sections focus on the traction and braking performance. Other performance, such as handling characteristics, noise, vibration and harshness (NVH), etc., will not be discussed in this entry.

Vehicle Performance

Vehicle performance discussed in this article will be restricted to propelling and braking performance in terms of vehicle speed, acceleration, gradeability, braking deceleration, and braking force distribution on front and rear wheels.

Overview of Vehicle Power Train Structure

Vehicle propelling performance is dictated by the power train structure, tractive power rating, and power plant operation characteristics and transmission design. Generally, vehicle power train consists of energy source (fuel, batteries), power plant (engine, electric motor), transmission, and drive wheels. The power train may be configured into front wheel drive, rear wheel drive, and all wheel drive, as shown in Fig. 1. The engine power is transmitted to drive wheels through clutch or torque converter, transmission, shaft, final drive, and transaxles. The torque on the drive wheels reacts with road surface to develop tractive effort to push the vehicle forward, as shown in Fig. 2.

The tractive effort developed on the drive wheel, together with the vehicle resistances, determines the vehicle performance (speed, acceleration, and gradeability).

Tractive Effort and Vehicle Speed

Tractive effort is produced on drive wheels by the reaction between the tractive torque and road surface as shown in Fig. 2. The tractive effort is proportional to the tractive torque, T_w, by

$$F_t = \frac{T_w}{r} \tag{1}$$

where T_w is the tractive torque on the drive wheels and r is the wheel radius.

The tractive torque acting on the drive wheels is produced by the engine or motor, transmitted through transmission gear box and final drive. It can be expressed as

$$T_w = T_e\, i_0\, i_g\, \eta_t \tag{2}$$

where T_e is the engine torque, i_0 is the gear ratio of the final drive, i_g is the gear ratio of the transmission, and η_t is the efficiency from the engine to the drive wheels. Combining (1) and (2), the tractive effort can be further expressed as

$$F_t = \frac{T_w\, i_0\, i_g \eta_t}{r} \tag{3}$$

The gear ratios of final drive and transmission are defined as the ratio of the input speed to the output speed. The transmission efficiency should include all the losses in the driveline components, such as, torque converter, transmission, torque distributor, final drive, etc. The typical efficiency values for individual components may be evaluated by [1, 2]

Clutch	99%
Each pair of meshed gears	95–97%
Bearing and joint	98–99%

It should be mentioned here that efficiency of torque converter is closely related to its operating point, that is, the speed ratio of the torque converter [1, 2].

Vehicle speed is related to the rotating speed of the drive wheels as

$$V = \omega_w\, r \quad \text{(m/s)} \tag{4}$$

where ω_w is the rotating speed of the drive wheels in rad/s. In terms of revolution per second (rpm) N_w, the vehicle speed can be further expressed as

$$V = \frac{\pi N_w\, r}{30} \quad \text{(m/s)}. \tag{5}$$

The wheel rotating speed is related to the engine rotating speed by

$$N_w = \frac{N_e}{i_0\, i_g}, \tag{6}$$

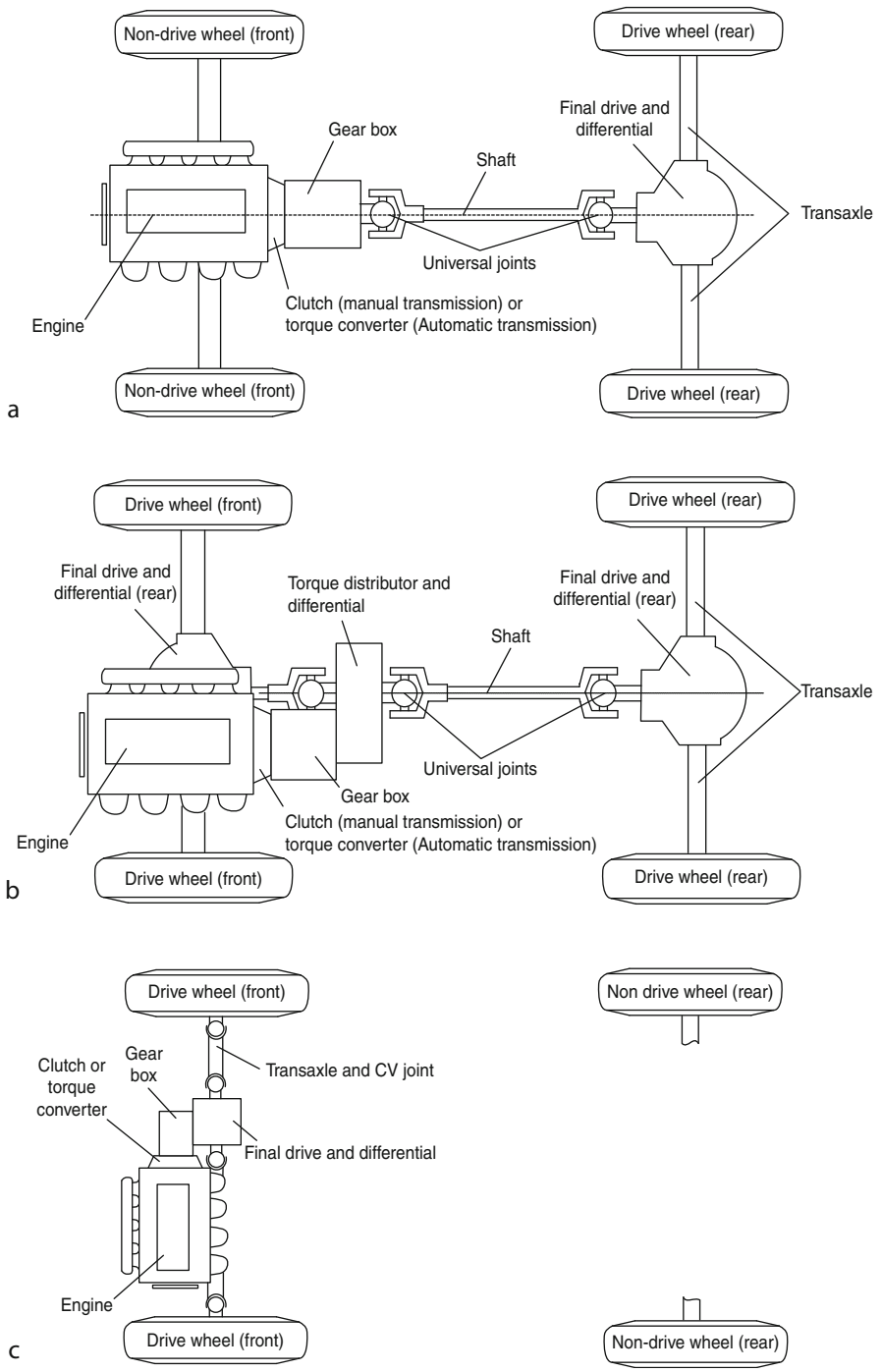
where N_e is engine rpm.

It should be noted that gear ratio of torque converter should be included into the gear box gear ratio, i_g. Gear ratio of torque converter is the reciprocal of the speed ratio of the torque converter [1, 2].

Combining (5) and (6), the vehicle speed can be further expressed as

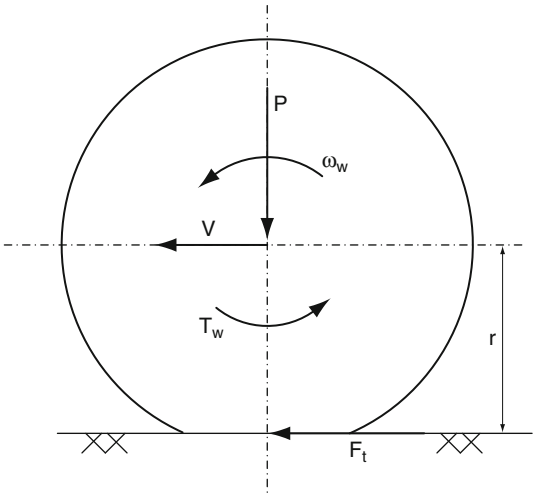
$$V = \frac{\pi N_e\, r}{30\, i_0\, i_g} \quad \text{(m/s)} \tag{7}$$

Equation 3 is only valid while the vehicle is running on well-prepared road, where no obvious slip



Vehicle Dynamics and Performance. Figure 1

Typical vehicle power trains: (**a**) rear wheel drive, (**b**) front wheel drive, and (**c**) all wheel drive



Vehicle Dynamics and Performance. Figure 2
Tractive effort, F_t, produced by wheel torque T_w

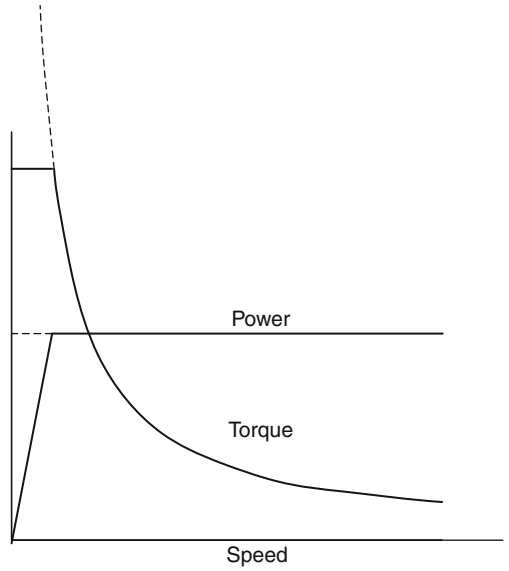
between tire and road surface occurs. For off-road operation, refer [1].

Torque (Power)-Speed Characteristics

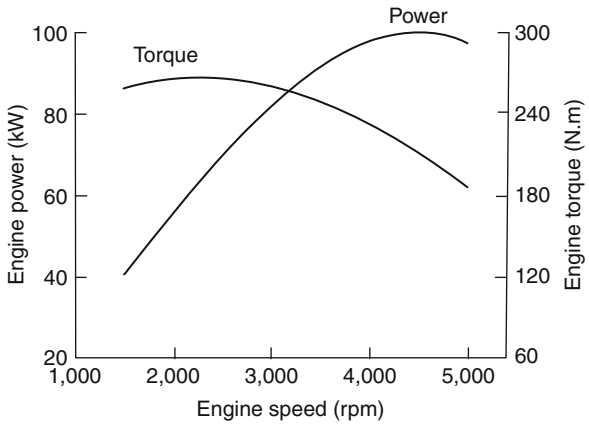
For maximizing vehicle performance at a given power plant power capacity, the ideal power plant is one that can produce a constant power in its full speed range. In this case, the power plant can apply its maximum power for propelling the vehicle in its whole speed range, yielding the highest vehicle performance. However, in practice, the maximum torque acting on drive wheels is limited by road adhesive capability. Beyond this limitation, obvious tire slip will occurs. Therefore, at low speed, high torque is usually cut off with a constant torque as shown in Fig. 3.

For evaluating vehicle performance, such as maximum speed, acceleration, and gradeability, the maximum tractive torques in its whole speed range are used, that is, the torque produced by a fully opened throttle engine. Figure 4 shows a typical gasoline engine operation characteristics with full open throttle [2]. Figure 5 shows the operation characteristics of a typical gasoline engine with partial open throttle.

The torque-speed profile of the engine is quit flat and is very different from the ideal profile shown in Fig. 3. Consequently, a multi-gear or continuous varying transmission (CVT) is required to modify this profile as shown in Fig. 6, in which, the envelope curve is a constant power curve as required.

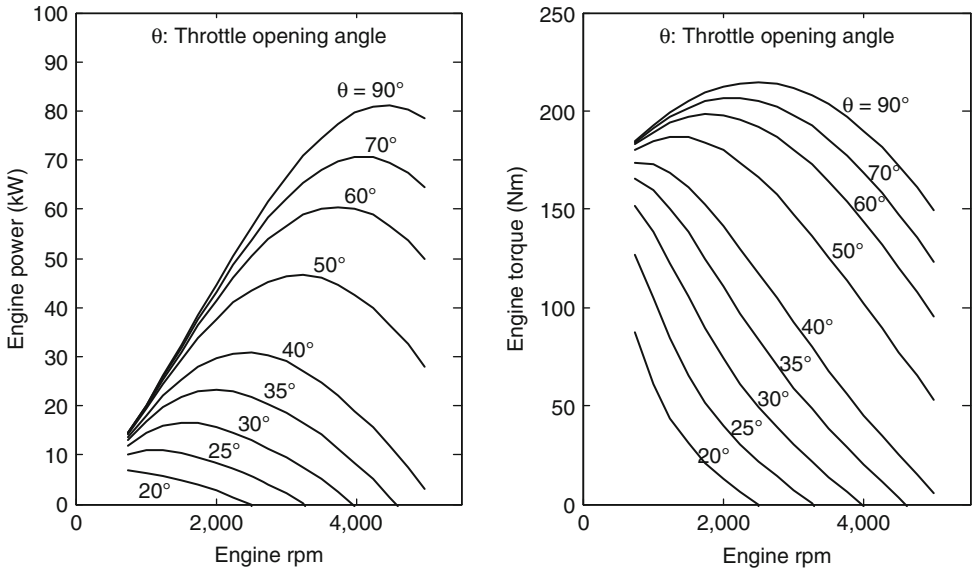


Vehicle Dynamics and Performance. Figure 3
Ideal torque (power)-speed performance of a vehicle power plant

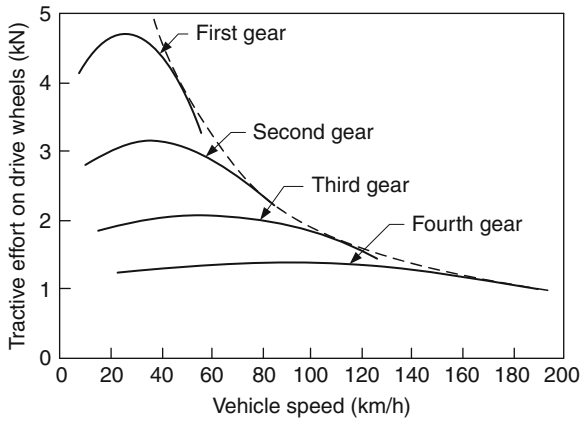


Vehicle Dynamics and Performance. Figure 4
Typical performance characteristics of gasoline engine with full open throttle [1]

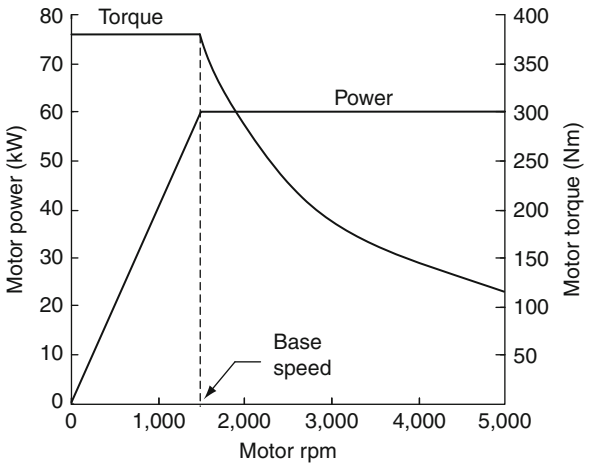
Compared with the toque-speed profile of internal combustion engine, well-controlled electric motors possess the torque-speed profile much closer to the ideal as shown in Fig. 7. At low speeds, the motor develops constant torque, and at high speeds, constant power. The corner speed is called as base speed. A traction motor is usually controlled in such ways



Vehicle Dynamics and Performance. Figure 5
Typical performance characteristics of gasoline engine with partial open throttle



Vehicle Dynamics and Performance. Figure 6
Gasoline engine–powered tractive effort with a multi-gear transmission



Vehicle Dynamics and Performance. Figure 7
Typical torque-speed profile of a well-controlled electric motor

that terminal voltage is linearly increased from zero speed to its base speed; meanwhile, the magnetic field is kept constant. Beyond the base speed, the terminal voltage is kept constant and magnetic field is linearly weakened. Since its torque-speed profile is naturally closer to the required, an electric motor–driven vehicle usually needs fewer gears than that of an engine-driven vehicle. A single- or double-gear transmission may meet the performance requirement.

Transmission Characteristics

As discussed above, torque-speed characteristic of internal combustion engine is far away from the ideal one and the engine cannot be directly connected to wheels. A multi-gear transmission or continuous varying transmission (CVT) is installed between the engine and drive wheels for modifying the torque-speed

characteristic of the engine output. A transmission is a gear box, in which there are several gear ratios for use. The gears are selected by driver based on real-time operation, such as vehicle speed and tractive effort requirement. This kind of transmission is referred to as manual transmission. The gear selection action may be performed automatically by a transmission control actuator, which is referred as automation or automatic transmission.

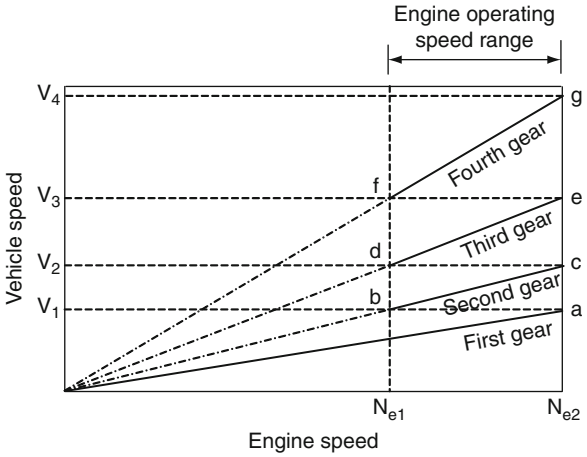
Transmission gear ratio design plays an importance role in vehicle performance. It should ensure the engine operating in its best speed range for torque producing and low fuel consumption.

Generally, there are several gears in a gear box. Each gear has a fixed gear ratio. Gear number depends upon vehicle performance requirements and specific power of the vehicle, which is defined as the power capacity per unit vehicle weight. More gear number results in a torque-speed profile closer to the ideal one, thus better vehicle performance can be obtained. However, more gears leads to complex transmission structure, and increased manufacturing difficulty, volume, weight, and cost. Therefore, the gear number design is usually a trade-off. In general, fewer gears are used in the vehicles that have large specific power. For passenger cars, four or five gears are common. For heavy-duty vehicles, over ten gears are common and two gear boxes are usually used.

Gear ratio design should ensure the engine operating in the best speed range. The design rule is interpreted as following.

As indicated by (7), for each gear, vehicle speed is proportional to engine speed. Gear changing procedure in acceleration driving is interpreted by Fig. 8 with a four-gear transmission.

1. Vehicle starts with first gear from zero speed to a speed, V_1 (point *a*), at which the engine speed reaches N_{e2}.
2. At point *a*, gear is changed from first to second (from point *a* to point *b*), in which vehicle speed does not change due to the very short gear change duration.
3. Start from point *b*, second gear is engaged, and continuous acceleration to point *c*.
4. At point *c*, gear is changed to third (from point *c* to point *d*).



Vehicle Dynamics and Performance. Figure 8
Engine speed versus vehicle speed at each gear

Using the speed relations, for example, the speed at point *a* equals to the speed at point *b*, the speed at point *c* equals to the speed at point *d*, and so on, and using (7), one obtains

$$V_1 = \frac{\pi N_{e2}\, r}{30\, i_0\, i_{g1}} = \frac{\pi N_{e1}\, r}{30\, i_0\, i_{g2}}$$
$$V_2 = \frac{\pi N_{e2}\, r}{30\, i_0\, i_{g2}} = \frac{\pi N_{e1}\, r}{30\, i_0\, i_{g3}} \qquad (8)$$
$$V_3 = \frac{\pi N_{e2}\, r}{30\, i_0\, i_{g3}} = \frac{\pi N_{e1}\, r}{30\, i_0\, i_{g4}}$$

Equation 8 gives

$$\frac{N_{e2}}{N_{e1}} = \frac{i_{g1}}{i_{g2}} = \frac{i_{g2}}{i_{g3}} = \ldots\ldots = \frac{i_{g,n-1}}{i_{g,n}} = k \qquad (9)$$

Equation 9 indicates that the ratios of two adjacent gears, *k*, is a constant, which determines the engine speed range, N_{e1}–N_{e2}. Small *k* allows engine operating in a narrow speed range, but more gears are needed.

When gear number and ratio, *k*, are determined, the gear ratio of each gear can be obtained using (9).

In transmission design, gear ratio of the highest speed gear is usually designed by selecting an engine speed, at which the vehicle has its top speed, that is,

$$i_{gh} = \frac{\pi\, N_{ed}\, r}{30\, i_0\, V_{\max}} \qquad (10)$$

where i_{gh} is the gear ratio of highest speed gear, $V_{\max}$ is the top speed of the vehicle with the highest speed gear,

N_{ed} is the desired engine speed with highest speed gear at maximum vehicle speed, and r is tire radius. After determination of the gear ratio of the highest speed gear, gear ratios of other gears can be determined by (9). It should be pointed out that the gear ratio of the lowest speed gear (first gear) should ensure the vehicle has its maximum tractive effort at low speed for meeting gradeability requirement.

In practice, gear ratio design may not exactly follow the rule of (9). For passenger cars, high-speed gears are used more often than low-speed gears in normal driving, thus gear ratios are usually designed more "dense" than low-speed gears, that is,

$$\frac{i_{g1}}{i_{g2}} > \frac{i_{g2}}{i_{g3}} > \ldots\ldots > \frac{i_{g,n-1}}{i_{g,n}}. \tag{11}$$

Vehicle Resistance

In operation, the tractive effort needs to overcome vehicle resistances and inertias. In steady-state operation, vehicle resistance consists of rolling resistance, aerodynamic drag, and gravity component along the vehicle moving direction during climbing a grade. In acceleration, the tractive effort also needs to overcome the vehicle inertias for picking up its speed.

Rolling Resistance Rolling resistance stems from the energy loss inside tire due to hysteresis effect of rubber materials and deformation of road surface [1, 2]. When vehicle is running on well-prepared hard road, the rolling resistance is mainly caused by tire hysteresis effect. However, when vehicle is running off-road, the ground deformation becomes the major factor.

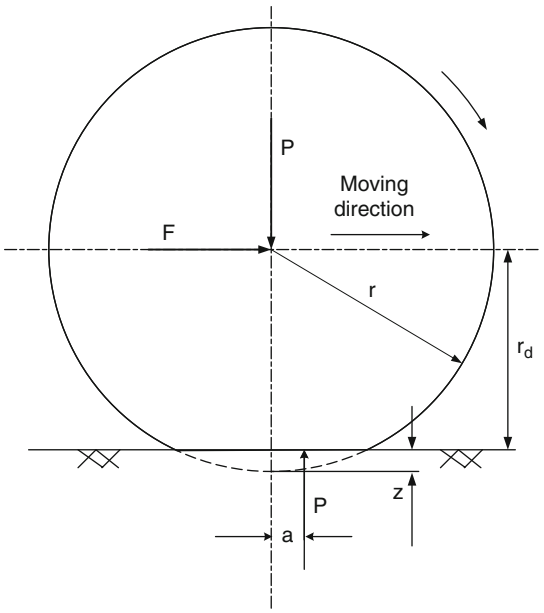
The rolling resistance is usually interpreted by a rolling coefficient, f_r defined as the horizontal force acting on the wheel rotating center, which maintains the wheel rotating on a road while unit load acting on the wheel center perpendicular to the road surface as shown in Fig. 9.

The rolling resistance of a vehicle is mathematically expressed as

$$F_r = M\, g\, f_r \cos\alpha, \tag{12}$$

where M is vehicle mass, g is gravity acceleration, 9.81 m/s^2, f_r is the rolling resistance coefficient, and α is the road grade angle as shown in Fig. 12.

The rolling coefficient is close to the tire and road conditions, such as tire material, tire structure, temperature, inflation pressure, tread geometry, road roughness, and presence of liquid on the road. In vehicle performance evaluation, at not very high speeds, rolling resistance can be taken as constant. Typical values on various roads are listed in Table 1 [1].



Vehicle Dynamics and Performance. Figure 9 Rolling resistance due to hysteresis effect inside tire material [2]

Vehicle Dynamics and Performance. Table 1 Rolling resistance coefficient [1]

Conditions	Rolling resistance coefficient
Car tire on a concrete or asphalt road	0.013
Car tire on a rolled gravel road	0.02
Tar macadam road	0.025
Unpaved road	0.05
Field	0.1–0.35
Truck tire on concrete or asphalt road	0.006–0.01
Wheel on iron rail	0.001–0.001

Vehicle speed should be taken in account for more accuracy at high speeds. Many experiments have been performed for determination of the effect of vehicle speed to the rolling resistance. Many empirical formulas have been proposed for calculation of rolling resistance coefficient on hard road. For example, the rolling resistance coefficient for a passenger car running on concrete road may be calculated by [1]

$$f_r = f_0 + f_s\left(\frac{V}{100}\right)^{2.5}, \tag{13}$$

where V is vehicle speed in km/h, and f_0 and f_s depend on inflation pressure of the tire [1]. For most common range of tire pressure, the rolling resistance coefficient of passenger cars on hard concrete road may be estimated by [1]

$$f_r = 0.01\left(1 + \frac{V}{160}\right). \tag{14}$$

This equation is effective to predict acceptable accuracy for speeds up to 128 km/h [1].

Aerodynamic Drag A moving vehicle in air subjects a force that resists the vehicle's motion. This force is referred to as aerodynamic drag. The aerodynamic drag is mainly resulted from the pressure difference between front and back of the vehicle as shown in Fig. 10. When the vehicle is moving, high-pressure areas are formed ahead of the vehicle. On the other hand, low-pressure areas are also formed behind the vehicle. The pressure difference between the high pressure in front of the vehicle and low pressure behind the vehicle results in a resultant force that tries to stop the vehicle. One effective approach to reduce the aerodynamic drag is to design the vehicle body shape for minimizing the areas of both high pressure and low pressure.

Aerodynamic drag of a vehicle in Newton is calculated by

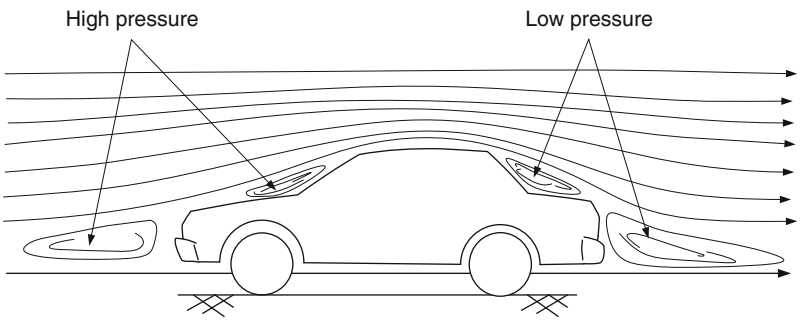
$$F_w = \frac{1}{2} C_D \rho_a A_f (V - V_w)^2, \tag{15}$$

where C_D is the aerodynamic drag coefficient that is determined by the vehicle body shape, ρ_a is the air density, 1.205 kg/m^3 for close earth surface, A_f is the front area of the vehicle body, V is the vehicle speed in m/s, and V_w is the wind velocity component in the vehicle moving direction, which has a positive value when this component in the same direction of the vehicle speed and negative when it is opposite to the vehicle speed. Typical values of aerodynamic drag for various types of vehicles are shown in Fig. 11 [3].

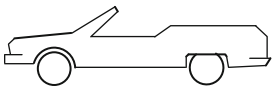
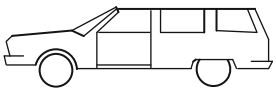
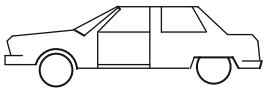
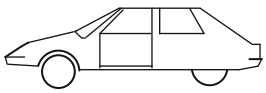
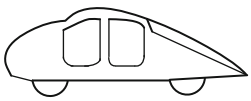
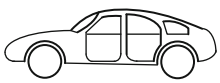
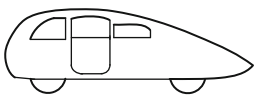
Grading Resistance The weight of a vehicle running on a grade produces a component that always directs toward downhill as shown in Fig. 12. This gravity weight component is a resistance for uphill and a drive force for downhill. For vehicle performance evaluation, only uphill operation is considered.

From Fig. 12, the grading resistance can be expressed as

$$F_i = M g \sin\alpha. \tag{16}$$



Vehicle Dynamics and Performance. Figure 10
Formation of aerodynamic drag

Vehicle type	Coefficient of aerodymanic drag
 Open convertible	0.5–0.7
 Van body	0.5–0.7
 Ponton body	0.4–0.55
 Wedged-shaped body; headlamps and bumper are integrated into the body, covered underbody, optimized cooling	0.3–0.4
 Headlamp and all wheels in body, covered underbody	0.2–0.25
 K-shaped (small reaksway section)	~ 0.23
 Optimum streamlined design	0.15...0.20
Trucks, road trains	0.8–1.5
Buses	0.6–0.7
Streamlined buses	0.3–0.4
Motorcycles	0.6–0.7

Vehicle Dynamics and Performance. Figure 11
Aerodynamic drag coefficients of various body shapes [3]

For simplifying calculation, $\sin\alpha$ is usually replaced by a grade value, i, with a small road angle. The road grade is defined as

$$i = \frac{H}{S} = \tan\alpha \approx \sin\alpha. \tag{17}$$

Further, the grading resistance can be expressed as

$$F_i = M\,g\,i. \tag{18}$$

Inertia Force When a vehicle is in acceleration, its translational and rotational inertias induce reaction force and reaction torque that are against the driving force and driving torque. For translational vehicle mass, the corresponding inertial force can be simply written as

$$F_{a-l} = M\,\frac{dV}{dt}. \tag{19}$$

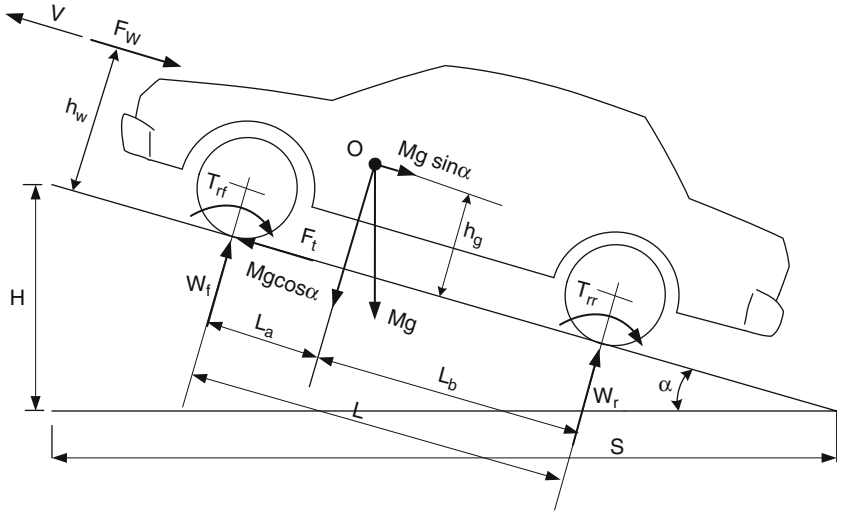
where dV/dt is the acceleration rate of the vehicle in m/s^2.

Along vehicle power line, there are many rotating components whose angular acceleration is related to the vehicle liner acceleration through gear ratios and wheel radius. For simplifying calculation, the rotational inertias are equivalently converted into translational inertia, which uses equal kinetic energy approach.

Suppose that a rotating component, i, in the drive line has the moment of inertia of J_i, and gear ratio of i_i to the drive wheel, using the equal kinetic energy principle, one obtains

$$\frac{1}{2}M_{i-e}V^2 = \frac{1}{2}J_i\omega_i^2. \tag{20}$$

where M_{i-e} is the equivalent linear mass inertia of rotating component i, V is the vehicle speed, and ω_i is



Vehicle Dynamics and Performance. Figure 12
Forces acting on uphill vehicle

the angular speed of the rotating component *i*. Thus, M_{i-e} can be expressed as

$$M_{i-e} = J_i \frac{\omega_i^2}{V^2}. \tag{21}$$

since ω_i and V have the relationship as

$$V = \frac{\omega_i r}{i_i} \tag{22}$$

Thus, (21) can be further expressed as

$$M_{i-e} = J_i \frac{i_i^2}{r^2} \tag{23}$$

where r is the radius of the drive wheel.

In vehicle acceleration performance analysis, only relative large rotating inertias are considered, typically the engine shaft components, such as flywheel, and running wheels. The equivalent translational mass inertia can be expressed as

$$M_{eq} = \frac{J_e i_0^2 i_g^2}{r^2} + \frac{J_w}{r^2} \tag{24}$$

where J_e is the moment of inertia of the rotating components that are attached on the engine shaft, i_0 and i_g are the final drive and transmission gear ratios, respectively, and J_w is the total moment of inertia of all wheels.

The total inertial force can be expressed as

$$F_a = (M + M_{eq})\frac{dV}{dt} = M(1 + \frac{M_{eq}}{M})\frac{dV}{dt} = M\delta\frac{dV}{dt} \tag{25}$$

where $\delta = (1 + M_{eq}/M)$ is defined as the equivalent mass factor.

Calculating equivalent mass factor, δ, needs to know the moments of inertia of all the rotating components. In the case of not knowing these values, δ of passenger cars would be estimated by empirical equation as

$$\delta = 1 + \delta_1 + \delta_2 i_0^2 i_g^2 \tag{26}$$

where δ_1 represents the term that is related to the moment of inertia of running wheels with a estimated value of 0.04 and δ_2 represents the term that is related to the moments of inertia of the engine-attached components with an estimated value of 0.0025 [2].

Vehicle Performance

Vehicle performance represented by maximum speed, gradeability, and acceleration is completely determined by the vehicle tractive effort developed by the engine or electric motor and the resistance in the vehicle motion direction. Figure 12 illustrates the forces acting on a vehicle, which is running uphill.

The forces acting on the vehicle in the vehicle moving direction are tractive effort, F_t, and resistances including rolling resistance, aerodynamic drag, grading resistance, and inertial force induced by acceleration. All these forces are always in a balanced state, which can be described by

$$F_t = F_r + F_w + F_i + F_a, \tag{27}$$

or more detail

$$\frac{T_e\, i_0\, i_g \eta_t}{r} = M\, g\, f_r \cos\alpha + \frac{1}{2} C_D \rho_a A_f V^2 + M\, g\, \sin\alpha + M\, \delta \frac{dV}{dt}. \tag{28}$$

When grade angle α is small, $\cos\alpha \approx 1$ and $\sin\alpha \approx \tan\alpha = i$. Equation 28 can further be written as

$$\frac{T_e\, i_0\, i_g \eta_t}{r} = M\, g\, f_r + \frac{1}{2} C_D \rho_a A_f V^2 + M\, g\, i + M\, \delta \frac{dV}{dt} \tag{29}$$

Equation 28 or 29 interprets the general operation behavior of a vehicle and is used to analyze vehicle performance. Depicting (29) using a tractive effort versus rolling resistance and aerodynamic drag on grade road is very helpful for vehicle performance analysis as shown in Figs. 13 and 14.

1. Maximum vehicle speed

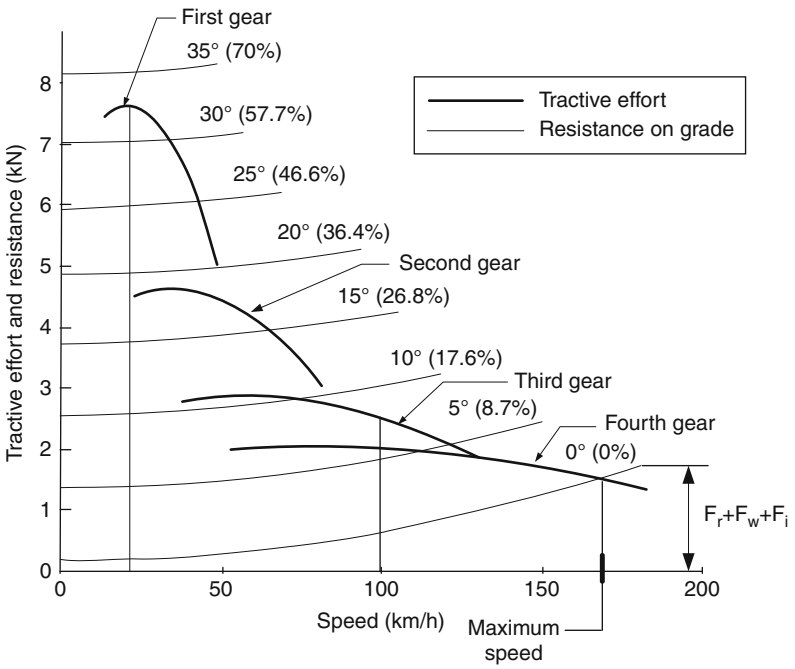
   The maximum speed of a vehicle is defined as the speed that can be reached when the power plant operates with its maximum capability (full open throttle for an IC engine and maximum current for an electric motor) on a flat road. Running on its maximum speed, no grading resistance and inertia force exists. Thus, (29) becomes

$$\frac{T_e\, i_0\, i_g \eta_t}{r} = M\, g\, f_r + \frac{1}{2} C_D \rho_a A_f V^2. \tag{30}$$

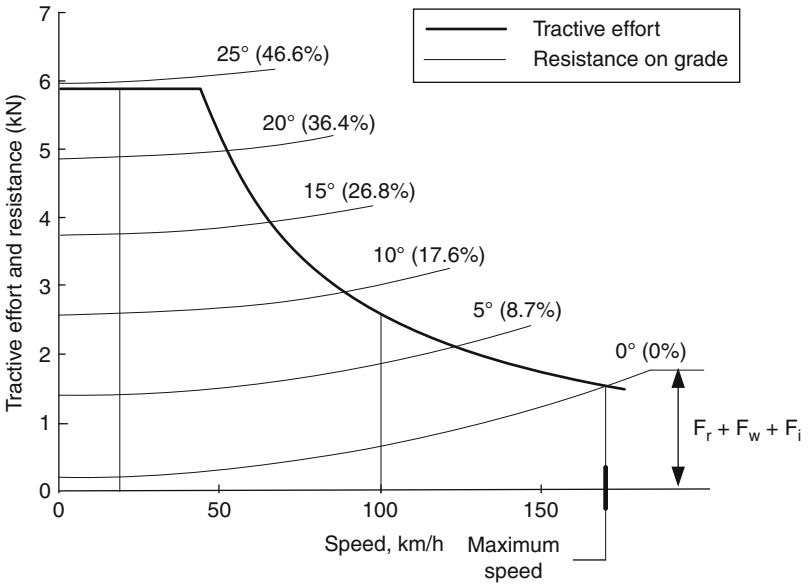
   On the diagrams of Figs. 13 and 14, maximum speed of vehicle can be obtained as the intersection of the tractive effort curve and the resistance curve at zero grade road.

2. Gradeability

   Gradeability of a vehicle is defined as the road grade or grade angle that the vehicle can overcome at a specified speed, for example, 100 km/h, or the



Vehicle Dynamics and Performance. Figure 13
Tractive effort versus vehicle resistance for a gasoline engine–powered vehicle [2]



Vehicle Dynamics and Performance. Figure 14
Tractive effort versus vehicle resistance for an electric motor–powered vehicle [2]

maximum grade at low speed. While running on an uphill grade with constant speed, 28 and 29 become

$$\frac{T_e i_0 i_g \eta_t}{r} = M g f_r \cos\alpha + \frac{1}{2} C_D \rho_a A_f V^2 + M g \sin\alpha \tag{31}$$

and

$$\frac{T_e i_0 i_g \eta_t}{r} = M g f_r + \frac{1}{2} C_D \rho_a A_f V^2 + M g i \tag{32}$$

The gradeability can be directly obtained by reading Figs. 13 and 14. For instance, for the gasoline engine–powered vehicle, running at 100 km/h, gradeabilities of 5.5° (9.6%) for fourth gear, 7.5° (13%) for third gear, and 32.5° (63.7%) at speed of about 20 km/h are obtained. Similarly, for the electric motor–powered vehicle, around 8° (14%) at the speed of 100 km/h and 24° (44.5%) at speeds lower than 50 km/h are obtained.

3. Acceleration performance

Acceleration performance of a vehicle is interpreted by the time used for accelerating the vehicle from a low speed (general zero speed) to a specified high speed (e.g., 100 km/h) on a level road. The acceleration time can be calculated by

$$t = \int_0^{V_f} \frac{1}{a} dV. \tag{33}$$

where V_f is the specified final speed and a is the acceleration rate in m/s^2, which can be obtained from (29) as

$$a = \frac{dV}{dt} = \frac{\frac{T_e i_0 i_g \eta_t}{r} - M g f_r + \frac{1}{2} C_D \rho_a A_f V^2}{M \delta} \tag{34}$$

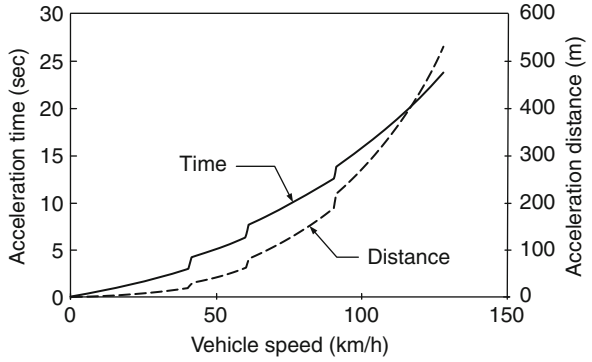
The distance covered during acceleration can be calculated by

$$d = \int_0^{t_a} V dt \tag{35}$$

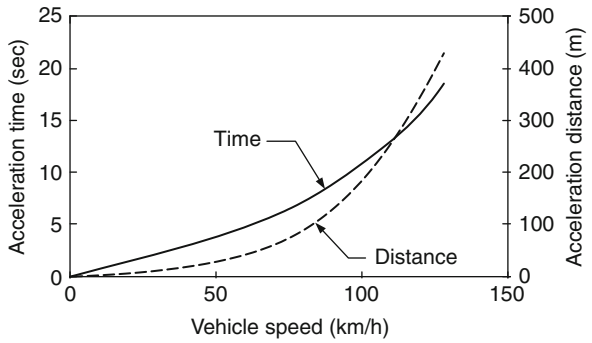
A digital integration method may be used to solve (33)–(35). Figures 15 and 16 show the acceleration time and distance for a gasoline engine–powered and electric motor–powered vehicle.

Fuel Economy and Energy Consumption

Fuel economy is one of the most important performances of a vehicle, which is evaluated by the fuel consumed in number of liters per 100 km. In the United States it is usually evaluated by the mileages per gallon fuel. In electric energy–driven vehicle, it is



Vehicle Dynamics and Performance. Figure 15
Acceleration time and distance of a gasoline engine–powered vehicle with a four-gear transmission



Vehicle Dynamics and Performance. Figure 16
Acceleration time and distance of an electric motor–powered vehicle with single-gear transmission

usually evaluated by kilowatt-hour (kWh) number of electric energy per unit traveling distance (km or mile). In plug-in hybrid vehicle, both fuel consumption and electric energy consumption are used at the same time.

Obviously, fuel and electric energy consumptions are associated with driving environments, typically, on highway and in urban. Highway driving is characterized by almost constant speeds. However, frequent stop-and-go is the common driving pattern in urban. These two very different driving patterns result in different fuel and energy consumptions. A vehicle specification usually lists the values for both highway and city driving.

Fuel and energy consumptions are determined by the energy consumed to overcome the vehicle resistances and the efficiency of the power train.

For engine-powered vehicle, the engine is the most inefficient component and its efficiency depends very much upon its operating points in terms of speed and torque or power. In practice, fuel usage efficiency of an engine is evaluated by fuel grams consumed per kWh energy output from its shaft, which is referred to as specific energy consumption (g/kWh). The typical fuel consumption characteristic of a gasoline engine is shown in Fig. 17 [2].

Fuel consumption varies with the operating points of the engine. The most efficient operating points are close to its full open throttle operation. At a given output power, the engine running at lower speed (point *a* in Fig. 17) results in lower fuel consumption than at higher speed (point *b* in Fig. 17). That is the reason that running in high-speed gear is more efficient than in low-speed gear. The transmission with more gears increases the chance for the engine operating close to its optimum operation line. Ideally, a continuous variable transmission (CVT) is capable of choosing a gear ratio that, at any driving condition, can operate the engine at its optimum line.

Based on the specific fuel consumption of engine, the fuel consumption rate (fuel amount consumed in unit time) of the engine can be determined by

$$Q_t = \frac{P_e\, g_e}{1000\gamma_f} \quad (\text{liter/h})\,, \tag{36}$$

where P_e is the engine power in kW, g_e is the specific fuel consumption (g/kWh) of the engine, which depends on its operating point (speed vs. power), and γ_f is the mass density of the fuel in kg/L. The engine power can be calculated by

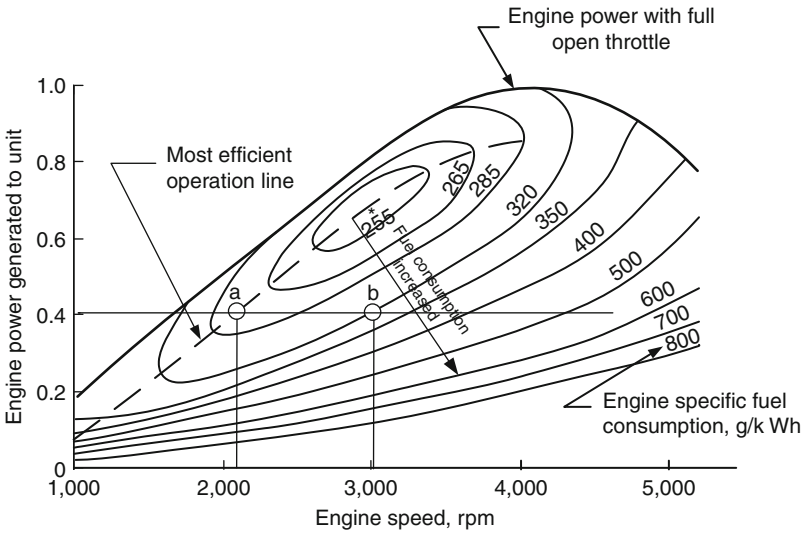
$$P_e = \frac{V}{\eta_t}(F_a + F_w + F_i + F_a). \tag{37}$$

Specific fuel consumption of the engine, g_e, can be obtained on the fuel consumption characteristic map (Fig. 17) with respect to the engine speeds, which can be determined by the vehicle speed and gear ratio as

$$N_e = \frac{30\, i_0\, i_g}{\pi\, r} V. \tag{38}$$

With constant speed, fuel consumption of a vehicle in a driving distance, S, is obtained by

$$Q_t = \frac{P_e\, g_e}{1000\gamma_f}\frac{S}{V}. \tag{39}$$



Vehicle Dynamics and Performance. Figure 17
Fuel consumption characteristics of a gasoline engine [2]

Since the complexity of vehicle operation in real world, fuel consumption at constant speeds cannot reflect the real fuel consumption scenario. Various driving cycles have been developed to emulate the real operation of a vehicle in specific driving environments. Figure 18 shows some typical driving cycles.

Driving cycles are interpreted by vehicle speed profiles versus driving time. For calculation of vehicle fuel consumption in a cycle, numerical computing method is usually used. The whole cycle is divided into many time intervals, usually 1 s, then, the fuel consumption in each time interval is calculated. The fuel consumption in the whole cycle can be obtained by summation of the fuel used in all the time intervals.

Figures 19 and 20 show the fuel consumption of a vehicle and engine operating points overlapping the engine fuel consumption map in FTP 75 urban and highway driving cycles [2]. It can be seen from these two diagrams that the engine operating points are far away from its most efficient area. This is one of the major reasons why conventional vehicle is inefficient, which hybrid technologies intend to cure.

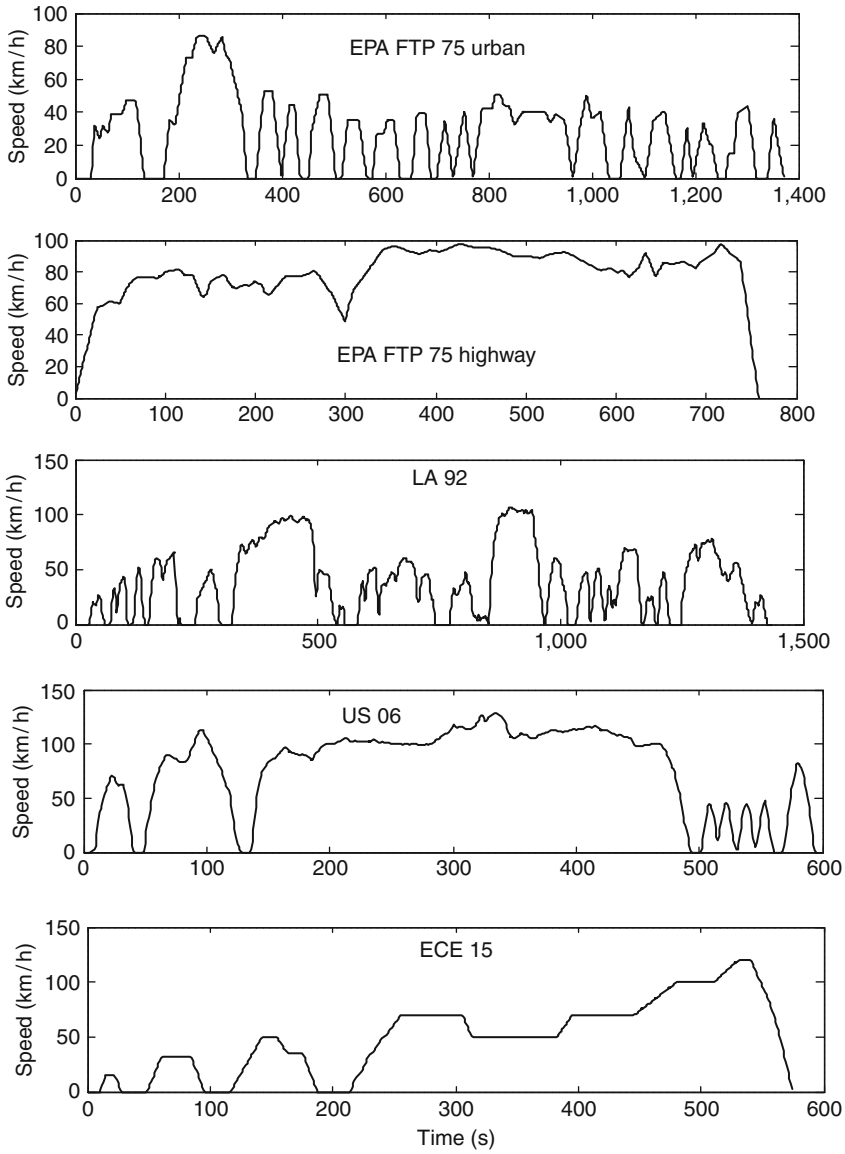
Vehicle Braking Performance

Braking performance of a vehicle is one of the most important requirements. The basic requirements of a vehicle brake system includes (1) fast reducing vehicle speed to complete stop, (2) braking effectiveness sustainability, and (3) maintaining vehicle stability during braking. Meanwhile, when a vehicle operates with stop-and-go pattern in urban areas, significant amount of energy is consumed, which is one of the major factors that result in low fuel efficiency. With more and more involvement of electric tractions, such as electric vehicle, hybrid electric vehicle, and fuel cell electric vehicle, braking energy recovery are becoming practice, in which, the kinetic energy of the vehicle body can be converted into electric energy by an electric machine. The recovered energy is charged into on-board energy storage, mostly chemical batteries, and can be reused in later traction. This is called regenerative braking. In vehicle brake system design, a new requirement is added, that is, recovering braking energy as much as possible. Nevertheless, braking safety is still the primary consideration.

Vehicle behavior during braking is determined by the reaction between the braking force and road. The basic requirements are to supply sufficient braking force to quickly stop the vehicle and at the same time, maintain the vehicle running direction stable and controllable.

Braking Force

Braking force of a vehicle determines the deceleration rate. Effective braking force is determined by two



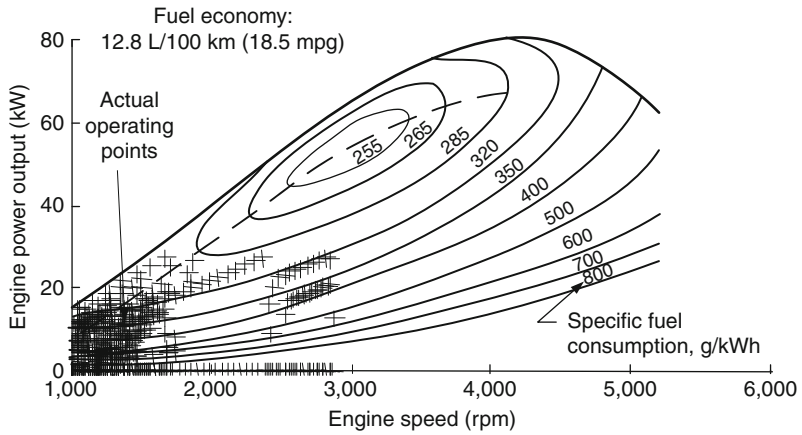
Vehicle Dynamics and Performance. Figure 18
Some typical driving cycles

factors. One is the braking torque developed by the brake system and the other is the road adhesive condition. When braking force is not large enough to lock the vehicle wheels, the braking force is determined by the braking torque generated by the braking system. However, when the wheels are locked, the braking force depends solely on the road adhesive condition.

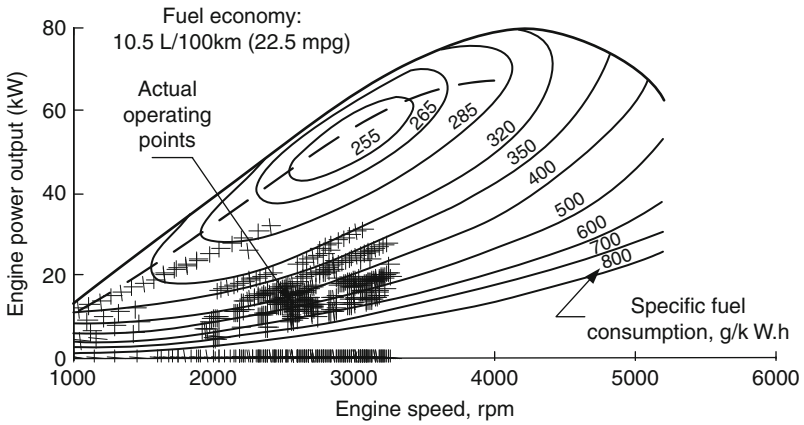
Figure 21 shows a braked wheel. The brake pad is pushed against the brake plate, generating a braking torque around the wheel's rotating center. This braking torque induces a braking force in the tire-ground contact area. In the case of unlocked wheel, the braking force is proportional to the braking torque as

$$F_b = \frac{T_b}{r}. \tag{40}$$

However, when the braking force goes up until locking the wheel, the braking force is completely



Vehicle Dynamics and Performance. Figure 19
Fuel consumption and engine operating points in EPA FTP 75 urban driving cycle [2]



Vehicle Dynamics and Performance. Figure 20
Fuel consumption and engine operating points in EPA FTP 75 highway driving cycle [2]

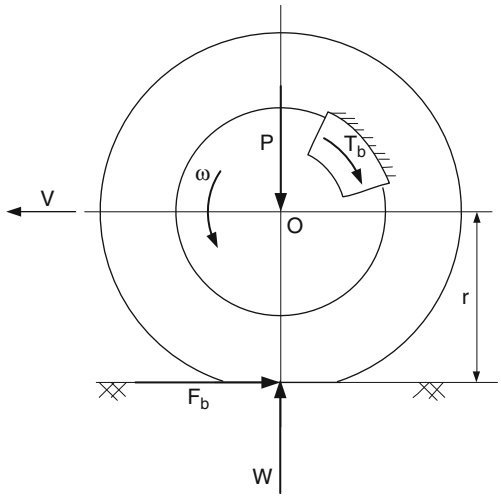
determined by the road adhesive capability, as shown in Fig. 22, which is expressed as

$$F_{b\max} = \mu W, \tag{41}$$

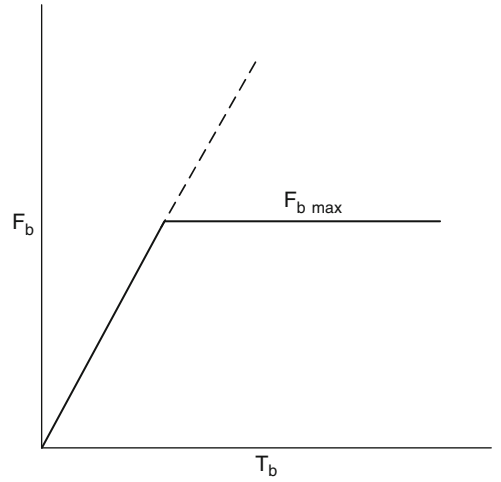
where W is the vertical load between the wheel and ground and μ is the adhesive coefficient of the wheel to ground. A well-prepared road has high adhesive coefficient. Contrary, a slippery road has low adhesive coefficient. Generally, adhesive coefficient is a function of wheel slip, s, as shown in Fig. 23. The wheel slip is defined as

$$s = \left(1 - \frac{r\omega}{V}\right) \times 100\%, \tag{42}$$

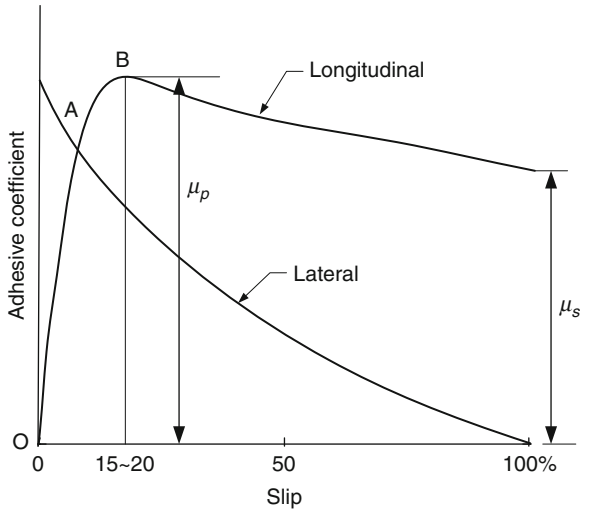
where V is vehicle speed (the translator speed of the wheel center), ω is angular speed of the wheel, and r is the wheel radius. With a free rotating wheel, $r\omega = V$ and $s = 0$. On the other hand, with a locked wheel, $\omega = 0$ and $s = 100\%$.



Vehicle Dynamics and Performance. Figure 21
Braking torque and braking force on a braked wheel



Vehicle Dynamics and Performance. Figure 22
Braking force versus braking torque



Vehicle Dynamics and Performance. Figure 23
Adhesive coefficient varying with wheel slip

Figure 23 indicates that with a small slip (between 0 and A), the adhesive coefficient is almost linearly proportional to the slip value. In this case, no obvious slip occurs in the wheel-ground area. The small slip is caused by the elastic property of the tire. Further increase in the braking force will cause actual slip at some points in the contact area. With increase in the braking force, more and more points go into slipping. In this case, the relationship between the slip and adhesive coefficient is nonlinear with a small increase rate. The adhesive coefficient reaches its maximum at a slip of about 15–20%. Beyond this peak point, the wheel will become unstable and is quickly locked up, even no further increase in the braking force. The adhesive coefficient at complete slipping is generally named as slipping value, which is smaller than the peak value.

It should be noted that the adhesive coefficient in lateral direction drops monotonously with the increase of wheel slip. At complete locked wheel, it becomes close to zero. Lateral adhesive coefficient represents the capability of resisting lateral disturbance. A completely locked wheel loses this capability, consequently, leading to unstable vehicle behavior during braking. Anti-lock brake systems (ABS) have been developed in order to prevent wheels from being completely locked, therefore improving vehicle braking stability.

Table 2 shows the average values of adhesive coefficients on various roads [1, 2].

Braking Strength and Braking Force Distribution on Front and Rear Wheels

Braking strength is defined as the deceleration rate, m/s^2 or g ($g = 9.81\ m/s^2$), which is completely determined by the total forces acting on the vehicle in the opposite direction of the vehicle travel. Rolling

resistance and aerodynamic drag also functions as braking forces. However, they are quit small, compared to road braking force, and ignored in braking performance analysis.

Figure 24 shows all the forces acting on a vehicle during braking on a flat road, where j is the braking strength in m/s^2, and F_{bf} and F_{br} are the braking forces acting on the front and rear wheels, respectively. The braking strength, j can be expressed as

$$j = \frac{F_{bf} + F_{br}}{M}, \tag{43}$$

where M is the vehicle mass.

Vehicle Dynamics and Performance. Table 2 Average values of adhesive coefficient on various roads [1, 2]

Surface	Peaking values, μ_p	Slipping values, μ_s
Asphalt and concrete (dry)	0.8–0.9	0.75
Concrete (wet)	0.8	0.7
Asphalt (wet)	0.5–0.7	0.45–0.6
Grave	0.6	0.55
Earth road (dry)	0.68	0.65
Earth road (wet)	0.55	0.4–0.5
Snow (hard packed)	0.2	0.15
Ice	0.1	0.07

It is considered to be ideal design for distribution of the total braking force on front and rear wheels in such a way that the slips on the front and rear wheels are always equal. This strategy can ensure front and rear wheels reach their maximum road adhesion at the same time, consequently, achieving the shortest braking distance. This approach is interpreted as that the braking forces are proportional to their vertical loads, which is expressed as

$$\frac{F_{bf}}{W_f} = \frac{F_{br}}{W_r} \tag{44}$$

Referring to Fig. 24, the vertical loads on front and rear wheels can be obtained as

$$W_f = \frac{M\,g}{L}\left(L_b + h_g\frac{j}{g}\right), \tag{45}$$

and,

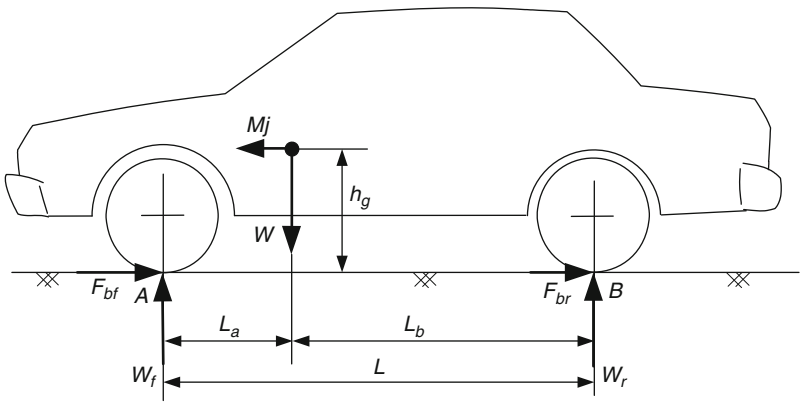
$$W_r = \frac{M\,g}{L}\left(L_b - h_g\frac{j}{g}\right). \tag{46}$$

Then, (44) can be further expressed as

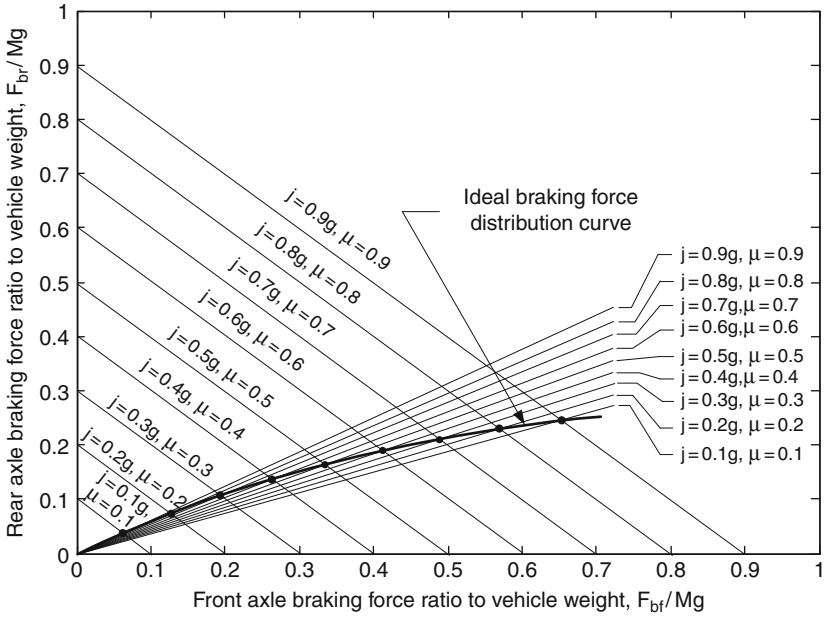
$$\frac{F_{bf}}{F_{br}} = \frac{L_b + h_g j/g}{L_a - h_g j/g}. \tag{47}$$

Equations 43 and 47 dictate the ideal braking force distribution on front and rear wheels, which can further be written as

$$F_{br} = \frac{M\,g}{2h_g}\sqrt{L_b^2 + \frac{4h_g L}{M\,g}F_{bf}} - \left(F_{bf} + \frac{M\,g\,L_b}{2h_g}\right) \tag{48}$$



Vehicle Dynamics and Performance. Figure 24
Forces acting on a vehicle during braking



Vehicle Dynamics and Performance. Figure 25
Ideal braking force distribution on front and rear wheels [2]

Figure 25 shows the ideal distribution curve (simply I curve), which has a parabolic shape. This figure also shows two sets of lines that represent (43) and (47) with respect to various braking strengths. When the braking forces reach their maximum values determined by the wheel-road adhesion, the maximum braking strength is achieved as

$$j_{\max} = \frac{F_{bf\,\max} + F_{br\,\max}}{M} = \frac{(W_f + W_r)\mu}{M} = g\mu. \tag{49}$$

Due to the nonlinear property of the ideal braking force distribution curve, design of a brake system to follow the ideal distribution curve becomes complicated. However, implementation of advanced electronic control may make this happen.

In practice, braking forces on front and rear wheels are usually designed to have a linear relationship. The proportion is represented by a ratio of the braking force on front wheel to the total braking force of the vehicle, that is,

$$\beta = \frac{F_{bf}}{F_b} = \frac{F_{bf}}{F_{bf} + F_{br}}. \tag{50}$$

β is determined by the brake system design, such as the diameters of front and rear wheel cylinders. For a given β, the applied braking forces on the front and rear wheels can be expressed as

$$F_{bf} = \beta F_b, \tag{51}$$

and

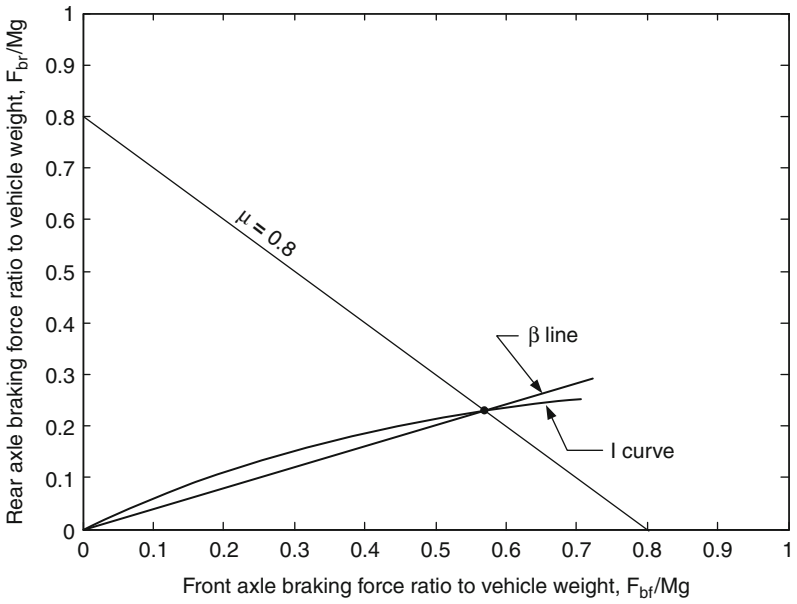
$$F_{br} = (1 - \beta)F_b, \tag{52}$$

and then

$$\frac{F_{bf}}{F_{br}} = \frac{\beta}{1 - \beta} \tag{53}$$

The braking force distribution represented by (53) is different from that described by (48) as shown in Fig. 26. One intersection point of these two lines exists, at which both front and rear wheels have the same slips, or both wheels are locked at the same time. This point specifies a specific adhesive coefficient, μ_0. Referring to (47), in which j/g is replaced by μ_0, one obtains

$$\frac{\beta}{1 - \beta} = \frac{L_b + h_g\mu_0}{L_a - h_g\mu_0}. \tag{54}$$



Vehicle Dynamics and Performance. Figure 26
Ideal and actual braking force distribution curve [2]

From (54), one obtains

$$\mu_0 = \frac{L\beta - L_b}{h_g} \tag{55}$$

As shown in Fig. 26, μ_0 divides the whole range into two sections. While braking on the road with an adhesive coefficient less than μ_0, front wheels are locked prior to rear wheels. Otherwise, rear wheels are locked prior to front wheels.

Braking Stability

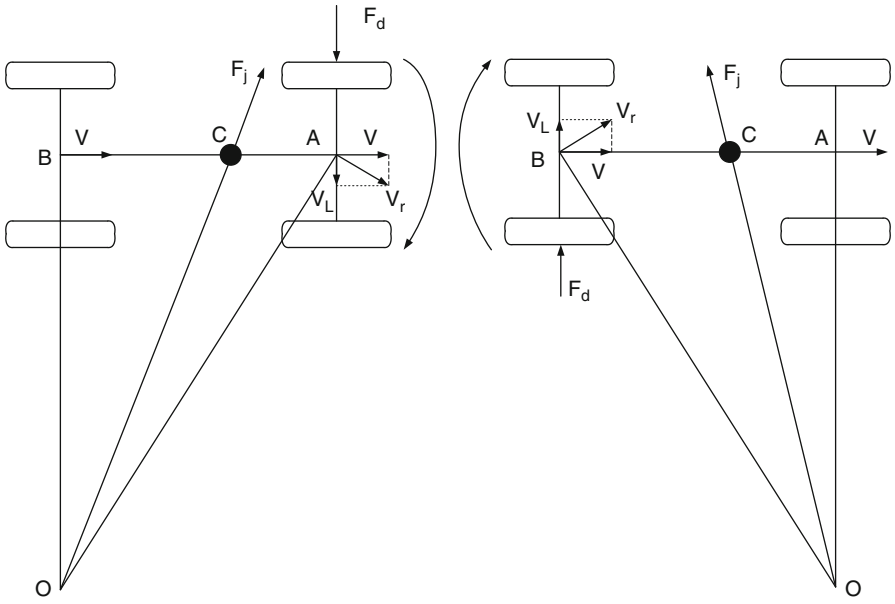
As discussed in a previous section, completely locked wheels will lose their capacity of resisting lateral disturbance. Consequently, any disturbance, such as wind, uneven road surface, and running along a curve road, would cause significant lateral movement and instability. Much more serious instability would occur with rear wheels locked than with front wheel locked. This can be interpreted in Fig. 27.

As shown in Fig. 27a, while a lateral slip on front wheels occurring, due to the lateral disturbance force F_d, the vehicle speed changes its motion from straightforward direction V to a curved running speed, V_r, which causes the whole vehicle body rotating around the point O. At the same time, the curved running of the vehicle body induces a centrifugal force, F_j, which acts on the gravity center, C, of the vehicle body. It can be seen from Fig. 27a that this centrifugal force has the effect of resisting against the lateral movement of the front wheel. When the disturbance force disappeared, the centrifugal force quickly brings the vehicle back to its straight motion.

While lateral slip occurs on rear wheels due to lateral disturbance, the induced centrifugal force augments the lateral slip, causing the vehicle very instable. Even after the lateral disturbance disappears, the lateral movement may continuous. Experiments have shown that when rear wheels are locked up over 0.5 s prior to the front wheels lockup, serious directional deviation would occur. In extreme case, vehicle body would swing 180° [1].

Front wheel lockup causes losing of steering capability. However, it can be detected more readily by the driver and correction can be made by the driver to release or partial release of the brake pedal. This is a less dangerous than rear wheel lockup swing.

Anti-lock brake system (ABS) can effectively present the wheels from lockup. This system monitors the wheel operation status. When it finds a wheel tending



Vehicle Dynamics and Performance. Figure 27
Vehicle behaviors with (**a**) lateral slip on front wheels and (**b**) lateral slip in rear wheels

to be locked, a control system reduces the braking force for this wheel and therefore brings the wheels back to its rotation.

Brake Design Regulation

For maintaining braking stability, rear wheel lockup is required not prior to the front wheel lockup. This requirement results in the braking force distribution always below ideal braking force distribution curve (I curve) as shown in Fig. 26. This can be realized by increasing the front wheel braking force and decreasing the rear wheel braking force. However, when most of the braking forces are applied to the front wheels and very small to rear wheels, it would cause a problem of reduced utilization of road adhesive capability. That is, when front wheels are locked and rear wheels are not locked, the maximum braking force of rear wheels is never used. Consequently, the stop distance will be longer. For avoiding this situation, some brake design regulations have been developed. A typical one is the ECE R13 regulation.

The ECE R13 regulation for passenger cars is represented by

$$\frac{F_{bf}}{W_f} \geq \frac{F_{br}}{W_r}, \tag{56}$$

and,

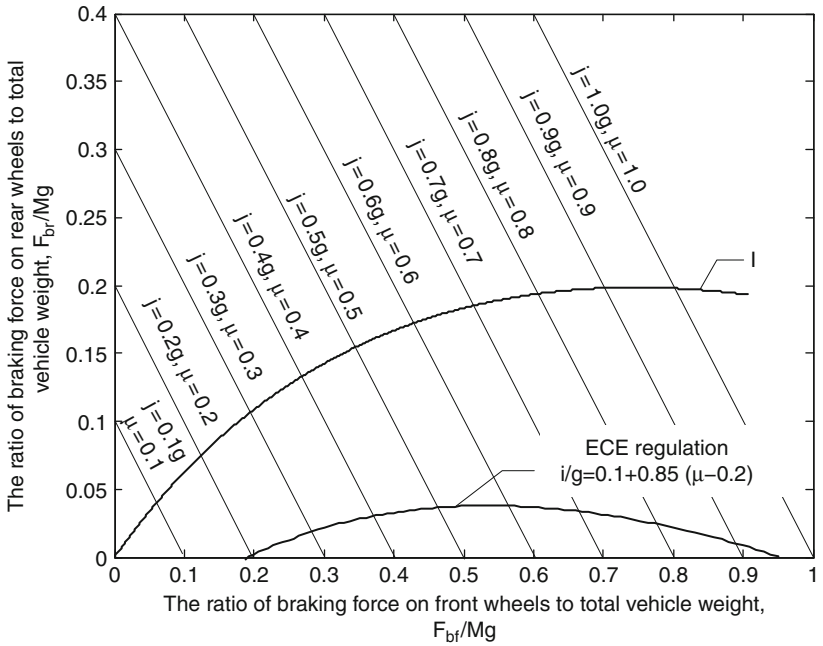
$$\frac{j}{g} \geq 0.1 + 0.85(\mu - 0.2). \tag{57}$$

Equation 56 dictates rear wheels being never locked prior to the front wheels, that is, β line is always below the I curve. Equation 57 dictates that, with locked front wheel, the braking force on rear wheels must be large enough to ensure to generate a braking strength j/g that is equal to or greater than a value dictated by (57).

Equations 56 and 57 are interpreted in a diagram as shown in Fig. 28. Obviously, the real braking force distribution design must fall into the area between the I curve and ECE regulation curve.

Future Directions

Vehicle system is a very complex system. With rapid development of computing technologies, the analysis methods of vehicle dynamics and performance are becoming more and more accurate and reliable by employing computer-based modeling and simulation. The advanced computing technologies allow solving much more complicated mathematical problems in



Vehicle Dynamics and Performance. Figure 28
The upper and low boundaries of braking force distribution dictated by ECE regulation

real time, which was impossible before computer age. Advanced data accessing, acquisition, processing, and control technologies provide great opportunity for vehicle dynamic control to improve vehicle performances.

Advanced electrical propulsion and electric energy storage create a way for high efficiency and clean vehicles that include electric vehicles, hybrid vehicles, and fuel cell vehicles.

Bibliography

Primary Literature

1. Wong JY (1978) Theory of ground vehicles. Wiley, New York
2. Ehsani M, Gao Y, Emadi A (2010) Modern electric, hybrid electric and fuel cell vehicles—fundamentals, theory and design, 2nd edn. CRC press, Boca Raton
3. Bosch R (2000) Automotive handbook. Tobert Bosch GmbH, Karsruhe

Books and Reviews

Gillespie TD (1992) Fundamentals of vehicle dynamics. SAE international, Warrendale
Mizutani S (1992) Car electronics. Sankaido Co., Tokyo

Vehicle Energy Storage: Batteries

Y. S. Wong[1], C. C. Chan[2]
[1]State Grid Energy Research Institute, State Grid Corporation of China, China
[2]Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China

Article Outline

Glossary

Battery A string of rechargeable electrochemical cells.
Battery electric vehicle An electric vehicle in which the electrical energy to drive the motor(s) is stored in an onboard battery.

Capacity The electrical charge that can be drawn from the battery before a specified cut-off voltage is reached.

Depth of discharge The ratio of discharged electrical charge to the rated capacity of a battery.

Electric vehicle A vehicle in which propulsion torque is delivered exclusively by one or more electric motors.

Energy capacity The electrical energy that can be drawn from the battery before a specified cut-off voltage is reached.

Fuel cell electric vehicle An electric vehicle in which the propulsion energy is delivered from an onboard fuel cell and battery hybrid system.

Hybrid electric vehicle A vehicle in which propulsion energy is provided from two or more kinds or types of energy stores, sources, or converters, and at least one of them delivers electrical energy.

Open circuit voltage The difference of electrical potential between two terminals of a battery when no external load is connected.

Vehicle energy source The onboard energy storage device of a vehicle.

Definition of the Subject

With ever-increasing concerns on energy efficiency, energy diversification, and environmental protection, electric vehicles (EVs), hybrid electric vehicles (HEVs), and low-emission vehicles are on the verge of commercialization. EVs not only offer higher energy efficiency than that of internal combustion engine (ICE) vehicles, but also mitigate one country's dependence on oil by diversifying the energy sources to renewable energies.

Vehicle energy source is bottleneck of EV and HEV commercialization. At present and in the foreseeable future, the viable energy sources for EVs and HEVs are batteries, fuel cells, and ultracapacitors (supercapacitors). The battery is the most mature energy source and it has been the most important component of an EV since commercialization of the first EV. This entry gives an overview of batteries for vehicle applications and discusses the research and development roadmap of next-generation batteries for vehicle applications.

Introduction

The EV has higher energy efficiency than that of the ICE vehicle and it also mitigates the one country's dependence on oil by diversifying the energy sources to renewable energies such as hydro, wind, and solar energies. The EV also facilitates load leveling of power systems and achieves zero local and minimal global vehicular emissions. At present, there is no economically viable energy source for commercialization of EVs.

The battery has been the most important component of an EV since commercialization of the first EV. In 1801, Richard Trevithick built a steam-powered carriage, opening the era of horseless transportation. The first battery-powered electric bicycle was built by Thomas Davenport in 1834. It was powered by a nonrechargeable battery and used on a short track. In 1838, Robert Davidson built a nonrechargeable battery-powered electric locomotive.

After the invention of lead-acid (Pb-Acid) battery in 1859, Sir David Salomons built a rechargeable battery-powered EV in 1874. The first petrol-powered ICE vehicle was built in 1885 and the first HEV was presented by J. Lohner and F. Porsche in 1901. The ICE vehicle outperformed the EV and HEV in the automotive century because there was no high performance battery for EVs and HEVs to overcome four major barriers to commercialization of EVs and HEVs, namely, short driving range, long charge time, long recharge time, and high life-cycle cost [1].

The battery for EVs has evolved from flooded lead acid (Pb-Acid) battery in 1859 to lithium ion (Li-Ion) polymer in 1999. Table 1 lists some historic events of vehicle batteries. The EV has also evolved a lot but it still holds its position in niche areas, like postal services [2].

The HEV has been introduced as an interim solution before the full implementation of the EV when there is a breakthrough in vehicle energy sources. The HEV extends greatly the driving range of the EV by three to four times and offers rapid fuel refueling. There are several types of EVs and HEVs in the market [3–5].

Technical requirements of batteries for vehicle applications are discussed by analyzing vehicle topologies and energy management systems in EVs' and HEVs' electrical powertrain. Viable batteries for EV and HEV applications are reviewed and the research and development roadmaps are discussed at the end of this entry.

Vehicle Energy Storage: Batteries. Table 1 Vehicle battery history

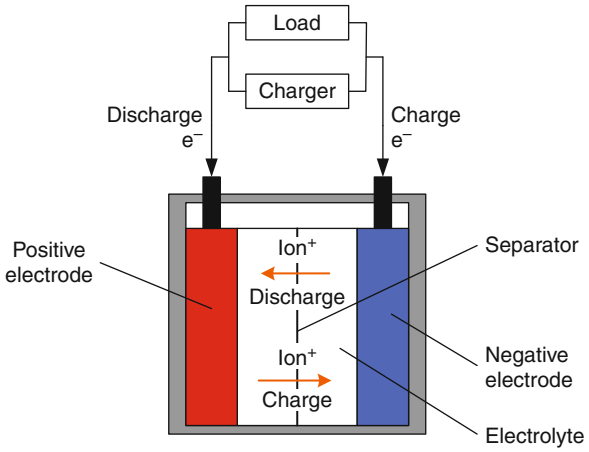
Year	Inventor	Battery
1859	Raymond Gaston Planté	Planté lead-acid cell
1881	Camille Alphonse Faure	Improved lead-acid cell
1899	Waldmer Junger	Nickel-cadmium cell
1899	Waldmer Junger	Nickel-iron cell
1946	Union Carbide Company	Alkaline manganese secondary cell
1970	Exxon laboratory	Lithium-titanium cell
1980	Moli Energy	Lithium-molybdenum disulfide
1990	Samsung	Nickel-metal hydride
1991	Sony	Lithium ion
1999	Sony	Lithium ion polymer

Electrical Powertrain

Motor drive and battery are two major components of an electrical powertrain in EVs and HEVs. Energy efficiency of motor drives is, intrinsically, higher than that of an ICE. In addition, the electrical powertrain can absorb kinetic energy during braking and assist engine in acceleration in HEVs. Thus, the efficiency improvement of an EV or an HEV over an ICE vehicle depends on power ratings and functionality of the electric motor drive and the onboard batteries [6, 7]. Key motor drive features and battery characteristics are discussed in the following sections.

Motor Drive Features

Regenerative Braking Regenerative braking is a unique feature of an electrical powertrain. The motor drive converts vehicle's kinetic energy back to electrical energy during braking, deceleration, and downhill running. The converted electrical energy is stored in the receptive energy sources such as batteries or ultracapacitors to extend the driving range. If the receptive sources are fully charged up, regenerative braking can no longer be applied and the vehicle is braked by the conventional hydraulic braking system.



Vehicle Energy Storage: Batteries. Figure 1 Basic components of an electrochemical cell

Power Boost The HEV allows both the engine and the electric motor to deliver power simultaneously to drive the vehicle. Generally, the engine and electric motor are coupled to the drive shaft of the wheel via two clutches, such that, the propulsion power may be supplied by the engine alone or by both of them. The motor cranks the engine and assists vehicle acceleration for maximization of engine fuel economy.

Battery Characteristics

Electrochemical Batteries The battery refers to the rechargeable electrochemical battery. Electrochemical cell is the basic element of each battery. A connection of a number of cells in series forms a battery. Figure 1 shows the basic components of an electrochemical cell in which both the positive and negative electrodes are immersed in the electrolyte. The electrolyte is an ion-conducting material, which can be in the form of aqueous, molten, or solid solution. The separator is a membrane that physically prevents direct contact between the two electrodes and allows ions, but not electrons, to pass through.

During discharge, the negative electrode performs oxidation reaction, which drives electrons to the external circuit, while the positive electrode carries out reduction reaction, which accepts electrons from the external circuit. During charge, the process is reversed so that electrons are injected into the negative electrode to perform reduction while the positive electrode releases electrons to carry out oxidation.

Battery Capacity Energy capacity (EC) of a battery refers to the electrical energy that can be drawn from the battery before a specified cut-off voltage is reached. EC is commonly presented in watt-hour (Wh). Coulometric capacity (C) refers to the total amount of electrical charge that can be drawn from the battery before the specified cut-off voltage is reached. C is typically measured in ampere-hour (Ah). C is widely employed to describe battery capacity, but EC is more common than C in addressing battery capacity in vehicle applications. Both EC and C, intrinsically, depend on the battery design, discharge current, temperature, and cyclic history.

Depth of Discharge and State of Charge Depth of discharge (DoD) and state of charge (SoC) are key parameters in battery energy management systems. DoD refers to the ratio of discharged electrical charge to the rated capacity. SoC refers to the ratio of usable charge to the rated battery capacity. SoC is an indicator of available electrical charge, while DoD is an indicator of discharged charge of a battery. Thus, sum of SoC and DoD of a battery is 100%. The SoC of a fully charged battery is 100% and the DoD is zero. When 20% of the stored charge is dissipated, SoC of this battery reduces to 80% and the DoD rises to 20% correspondingly.

Electrical Efficiency Energy efficiency of a battery refers the ratio of output electrical energy during discharging to the input electrical energy during charging. The energy efficiency is different from the charge efficiency, which is defined as the ratio of discharged charge to the charged charge. For vehicle applications, the energy efficiency is more informative than the charge efficiency. Typical energy efficiency of a battery is 55–75%.

Energy Density and Specific Energy Energy densities of an energy source refer to the usable energy capacity per unit mass or volume. The gravimetric energy density is usually named as the specific energy in watt-hour per kilogram (Wh/kg). The volumetric one is loosely named as energy density in watt-hour per litre (Wh/L). Specific energy is more instructive than the energy density for vehicle batteries because the battery weight is highly correlated with the vehicle fuel economy while the volume only affects the usable space. The specific energy is a key parameter to assess the pure electric driving range. The usable energy capacity greatly varies with discharge rate. The larger the discharge rate, the smaller the usable energy. Generally, specific energy and energy density are quoted with a discharge rate.

Power Density and Specific Power Power densities of a battery denote the deliverable rate of energy per unit mass or volume. The gravimetric power density is named specific power and measured in watt per kilogram (W/kg) and the volumetric power density is named power density and measured in watt per litre (W/L). The specific power changes with DoD. Thus, specific power and power density are quoted with DoD.

Cycle Life Cycle life is a key parameter to describe the service life of a battery based on the storage capacity of the battery. It is defined as the number of charge and discharge cycles it can undergo before its capability falls to 80% of the rated capacity. Cycle life is greatly affected by DoD of each discharge cycle, thus, it is usually quoted with DoD. For example, a battery can be claimed to offer 500 cycles at 80% DoD and 1,000 cycles at 50% DoD [1].

Calendar Life Calendar life refers to life period of a battery until failure in years. The battery calendar life depends on charge rate, discharge rate, DoD, temperature, and chemistries of the battery.

Power and Energy Demands of EVs and HEVs

EVs and HEVs can be further divided into six types of vehicles according to the demands of energy and power on vehicle batteries. Instead of grouping HEVs by vehicle architecture, it is more informative to group them by functionality of the electrical powertrain, which affects the fuel economy significantly.

HEVs are classified into four specific hybrids: micro hybrid vehicle (MHV), mild hybrid electric vehicle (MHV), full hybrid electric vehicle (FHEV), and plug-in hybrid electric vehicle (PHEV). On the other hand, EVs are classified into battery EV (BEV) and fuel cell EV (FCEV). A BEV is an EV where the electrical energy to drive the motor(s) is stored in onboard rechargeable batteries while an FCEV is an EV making

use of fuel cell and battery hybrid system as onboard energy sources [4, 8].

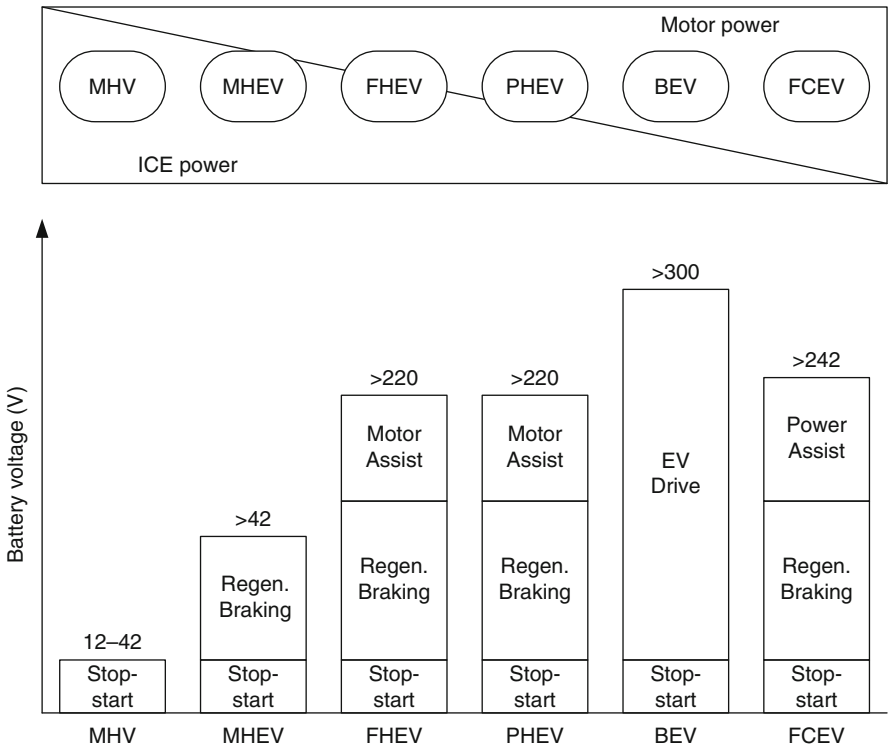
The battery is still the most important component of these vehicles but the requirements on power, energy, cycle life, and system voltage are different. Functionality of the electrical powertrain and the favorable battery voltages in these vehicles are shown in Fig. 2.

Hybrid Electric Vehicles

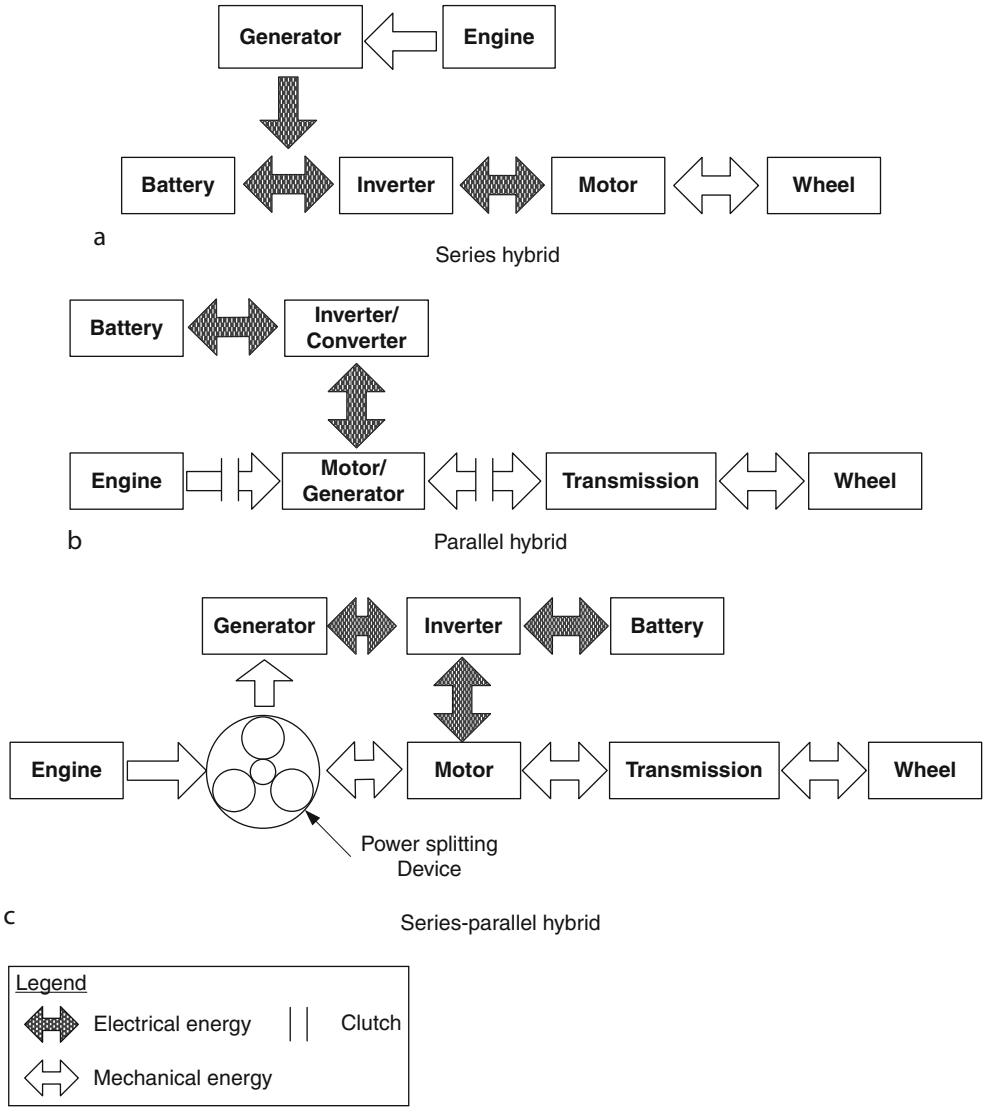
Energy management strategies of HEVs aim to satisfy four key goals: maximum fuel economy, minimum emissions, minimum system costs, and high acceleration rate. The major challenges of HEV design are management of multiple energy sources, battery sizing, and battery management. HEVs take the advantages of electric drive to compensate the inherent weakness of engine. HEVs can avoid engine idling and increase the engine efficiency during starting, low-speed, and high-speed operations. HEVs can also absorb energy during regenerative braking [3, 9].

Hybrid Electric Vehicle Configuration The three basic hybrid architectures of HEVs are series, parallel, and series-parallel hybrids. Figure 3 shows the energy paths in these hybrids. The series hybrid couples the engine and the battery by a generator. Both the engine and the battery power the electric motor to propel the vehicle. The parallel hybrid couples the mechanical power from the engine and from the electric motor to propel the vehicle. The series-parallel hybrid is a direct combination of the series and parallel hybrids.

Micro Hybrid Vehicles The MHV has an electric motor with peak power of about 2.5 kW. The electrical powertrain is driven by a battery system at 12–42 V. The motor is small and simple in structure. It can be an integration of starter and alternator in an ICE vehicle. The electrical and mechanical powertrains in an MHV are governed by an automatic stop-start mechanism, in which, the engine shuts down under vehicle braking and rest. The MHV is favorable for city driving, where



Vehicle Energy Storage: Batteries. Figure 2
Battery operating voltages in EVs and HEVs



Vehicle Energy Storage: Batteries. Figure 3
Energy flow in series, parallel, and series-parallel HEVs

there are frequent stops and starts. An MHV's fuel economy can be 5–10% higher than that of an ICE vehicle in city driving. The Citroen C3 is an MHV using the Valeo motor system.

The battery discharges frequently in cranking the engine in MHVs. Thus, there is a demand for high cycle life for batteries in MHVs. Table 2 lists some key technical data of batteries for MHVs.

Mild Hybrid Electric Vehicles The MHEV has a more powerful electrical powertrain than an MHV's. The typical electric motor power of a sedan MHEV is about 10–20 kW at 100–200 V. The motor is directly coupled with the engine. The motor has a large inertia such that it can replace the original flywheel of the engine. The motor and the engine are generally coupled in parallel hybrid configuration. Table 3 shows some technical data of batteries for MHEVs. The electrical powertrain is designed to crank the engine and perform regenerative braking during braking. There are demands of high specific power and long service life for batteries in MHEVs. Battery's charge and

Vehicle Energy Storage: Batteries. Table 2 Technical data of batteries for MHVs

Parameters	Unit	MHV
Voltage	V	12–42
Discharge power	kW	4.2–6
Low temperature (−28°C) discharge power	kW	>3
Energy capacity	kWh	0.2–1
Operating temperature	°C	−30 to +52
Calendar life	Year	>3

Vehicle Energy Storage: Batteries. Table 3 Technical data of batteries for MHEVs

Parameters	Unit	MHEV
Voltage	V	42–200
Discharge power	kW	>15
Low temperature (−28°C) discharge power	kW	>4
Recharge power	kW	>15
SoC window	%	40–70
Recharge pulse power	kW	>20
Energy capacity	kWh	0.8–1
Operating temperature	°C	−30 to +52
Calendar life	Year	>10

discharge power depend on its SoC. The battery's discharge power decreases with its SoC. The minimum operating SoC is around 40–50% to uphold sufficient power for launch and acceleration support. On the other hand, the battery's recharging power drops when the SoC is high, thus, the maximum operating SoC is regulated at around 70–80% to maintain sufficient recharge power for regenerative braking. Typically, the batteries operate in an SoC window between 40% and 70%.

Comparing with an ICE vehicle, the MHEV can boost the fuel economy by 20–30% in city driving. MHEVs in the market include Honda Insight Hybrid, Honda Civic Hybrid, and Ford Escape Hybrid.

Vehicle Energy Storage: Batteries. Table 4 Technical data of batteries for FHEVs

Parameters	Unit	FHEV
Voltage	V	220–350
Discharge power	kW	>35
Low temperature (−28°C) discharge power	kW	>4
Recharge power	kW	>30
SoC window	%	40–80
Recharge pulse power	kW	>40
Energy capacity	kWh	1–2
Operating temperature	°C	−30 to +52
Calendar life	Year	>10

Full Hybrid Electric Vehicles The FHEV has a high power electrical powertrain to drive the vehicle purely by electricity in a short driving range. The typical electric motor power for sedan FHEV is about 50 kW at 200–350 V. Generally, the motor, generator, and engine are coupled in series-parallel configuration. With the aid of power split devices, which are mainly built by planetary gear sets and clutches, the energy management system of the engine, motor, and generator is designed to maximize energy efficiency and minimize emissions.

The FHEV can be driven in pure EV mode and hybrid mode. The electrical powertrain assists the engine, not only at the starting, but also during acceleration in the hybrid model, which is also called charge-sustaining mode. In this mode, the discharged energy of the battery is recharged not only during braking but also by the engine to maintain the SoC in high and narrow window. Table 4 shows the technical data of batteries for FHEVs.

The FHEV can achieve higher fuel economy than that of the ICE vehicle by 30–50% in city driving. FHEVs in the market include Toyota Prius, Highlander, and Lexus RX 400 h.

Plug-in Hybrid Electric Vehicles The electrical powertrain of a PHEV is similar to that of an FHEV. The key differences are the additional battery pack and the functionality of grid recharging. In addition to the charge-sustaining mode, the PHEV can also operate in

the charge depletion mode, in which the PHEV operates in pure EV mode. Thus, the battery SoC drops in the charge depletion mode.

The electrical drivetrain of a PHEV works in a high voltage at 220–350 V. The battery energy capacity in PHEVs is the largest among all HEVs and it is determined by the targeted pure electric driving range. The PHEV operates in the charge depletion mode first and then the charge-sustaining mode. In the charge depletion mode, the batteries decline from 100% to a threshold SoC, which triggers the operation mode change. In the charge-sustaining mode, the battery SoC is regulated between the bottom of the SoC window and the threshold SoC. The battery is recharged from the grid at the end of the trip. Similar to the EV, the PHEV suffers from complexity and costliness. However, the PHEV delivers longer driving range than the EV's. Table 5 shows technical data of batteries for PHEVs.

The BYD F3DM is the world's first mass production PHEV, which went on sale to the government agencies and corporations in China in December 2008. Toyota also works on a plug-in version of the Prius. The plug-in Prius is converted from the Prius by adding additional 1.3 kWh battery pack into the car and a charging unit. The plug-in Prius and F3DM adopt the series-parallel hybrid powertrain.

Vehicle Energy Storage: Batteries. Table 5 Technical data of batteries for PHEVs

Parameters	Unit	PHEV
Voltage	V	220–350
Discharge power	kW	>50
Low temperature (−28°C) discharge power	kW	>6
Recharge power	kW	>30
SoC window	%	20–100
Recharge pulse power	kW	>20
Energy capacity	kWh	5–20
Charge time	Hour	<5
Operating temperature	°C	−30 to +52
Calendar life	Year	>10

A PHEV can also be implemented in a series hybrid topology. The GM Chevrolet Volt is a series PHEV, which is also called extended-range electric vehicle (EREV). The EREV is driven by one sole electrical powertrain, powered by the battery and a small engine.

Hybrid Electric Buses The battery in hybrid electric buses (HEBs) functions for engine start, power boost, and regenerative braking, which are very similar to the application in FHEVs. The HEB is heavier than a sedan HEV, such that the batteries needed for HEBs are a scale up of that in a sedan FHEV. The battery operates at a significantly higher voltage of 400–700 V. Table 6 shows the technical data of batteries for HEBs.

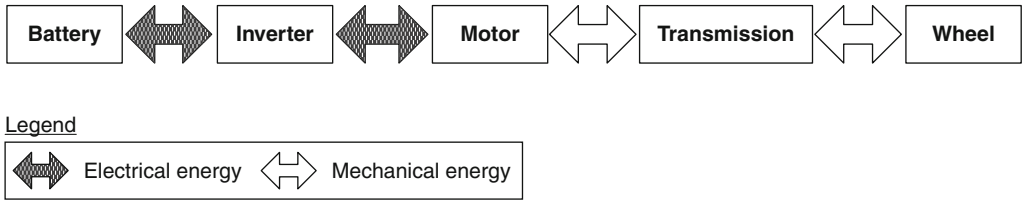
Electric Vehicles

Battery Electric Vehicles Battery is the sole energy source for the electrical powertrain and accessory systems in a BEV. The typical electric motor power for sedan BEV is about 50–80 kW and the battery operates in a high voltage, over 300 V. Figure 4 shows the bidirectional energy flow in the BEV powertrain and Table 7 lists some key data of batteries for BEVs. In city driving, the EV's fuel economy can be double that of an ICE vehicle [10, 11]. The Mitsubishi i-MiEV and Nissan Leaf are typical BEVs in the market.

Fuel Cell Electric Vehicles An FCEV is driven by a fuel cell and battery hybrid system. A fuel cell system

Vehicle Energy Storage: Batteries. Table 6 Technical data of batteries for HEBs

Parameters	Unit	HEB
Voltage	V	400–700
Discharge power	kW	100–200
Low temperature (−28°C) discharge power	kW	>20
Recharge power	kW	50–100
SoC window	%	40–70
Energy capacity	kWh	>10
Operating temperature	°C	−30 to +52
Calendar life	Year	>5



Vehicle Energy Storage: Batteries. Figure 4
Energy flow in a BEV

Vehicle Energy Storage: Batteries. Table 7 Technical data of batteries for BEVs

Parameters	Units	BEV
Voltage	V	>300
Discharge power	kW	>50
Low temperature (−28°C) discharge power	kW	>40
Recharge power	kW	>30
SoC window	%	20–100
Energy capaciy	kWh	50–90
Charge time	Hour	<8
Operating temperature	°C	−30 to +52
Calendar life	Year	>10

consists of a fuel cell stack with plant components for air supply, fuel control, temperature control, and humidification control. Most FCEVs combine the fuel cell stack with a high-voltage battery in a hybrid system for regenerative braking and acceleration boost. The batteries operate in an SoC window between 40% and 70%. Figure 5 shows the energy flow in an FCEV and Table 8 shows technical data of batteries for FCEVs [5].

Batteries for Vehicle Applications

The battery has to be intrinsically tolerant to abuse conditions such as overcharge, short circuit, crush, fire exposure, and mechanical shock and vibration. Battery cells are connected in series and parallel in a vehicle battery system, thus, cells' SoCs have to be balanced to prevent undercharge and overcharge. The operating temperature range of the battery is wide such that active thermal management systems are needed [1, 12].

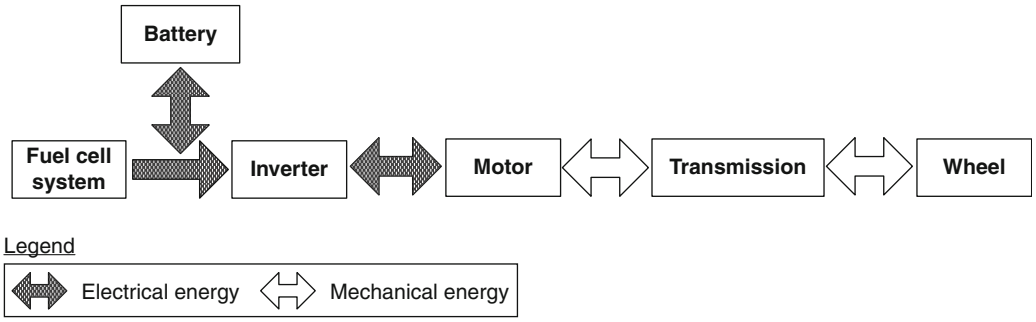
Key requirements for vehicle batteries are high specific energy and specific power, long cycle life, high efficiency, wide operating temperature, and low cost for commercialization. Figure 6 shows the power and energy requirements of battery for various EVs and HEVs.

The United States Council for Automotive Research LLC (USCAR) and the United States Advanced Battery Consortium (USABC) have set technical targets for vehicle batteries. The MHVs need batteries with long cycle life. The MHEVs, FCEVs, FHEVs, and HEBs need batteries with high specific power for power boost and regenerative braking. Table 9 shows some USABC's key goals for batteries in HEV applications [12–15].

The EV needs batteries with high specific power for quick charge and with high specific energy for long driving range. Table 10 shows some USABC's key goals for batteries in EV applications. The EV commercialization goals were developed to provide lower and possibly reachable goals for car manufacturers to enter the EV market in the near future.

The PHEV needs batteries with high specific power but the requirements on specific energy vary with the targeted pure electric driving range. Table 11 shows some USABC's key goals for batteries in PHEV applications. The high power to energy ratio battery is required for PHEVs with 10-mile pure electric driving range, while the high energy to power ratio battery is required for a pure electric driving range of 40 miles.

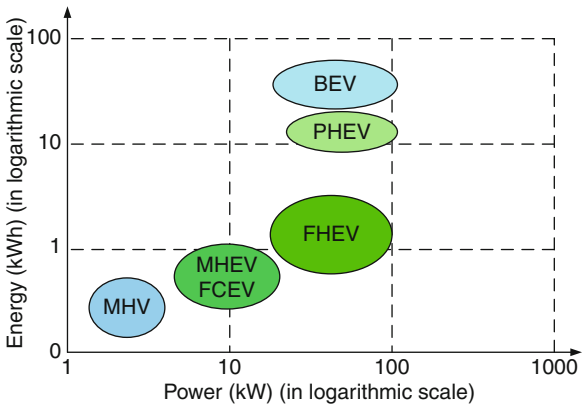
The viable batteries for vehicle applications consist of the valve-regulated lead-acid (VRLA), nickel-cadmium (Ni-Cd), nickel-zinc (Ni-Zn), nickel-metal hydride (Ni-MH), zinc/air (Zn/Air), aluminium/air (Al/Air), sodium/sulfur (Na/S), sodium/nickel chloride ($Na/NiCl_2$), lithium metal-polymer (LiM-Polymer), and lithium-ion (Li-Ion) batteries. The specific energy and specific power of these batteries are shown in Fig. 7. These batteries are classified into lead-acid,



Vehicle Energy Storage: Batteries. Figure 5
Energy flow in an FCEV

Vehicle Energy Storage: Batteries. Table 8 Typical specification of batteries for FCEVs

Parameters	Unit	FCEV
Voltage	V	200–440
Discharge power	kW	>15
Low temperature (−28°C) discharge power	kW	>4
Recharge power	kW	>15
SoC window	%	40–70
Recharge pulse power	kW	>20
Energy capacity	kWh	0.6–1.5
Operating temperature	°C	−30 to +52
Calendar life	Year	>10



Vehicle Energy Storage: Batteries. Figure 6
Power and energy requirements of batteries for various EVs and HEVs

nickel-based, zinc/halogen, metal/air, sodium-beta, and ambient-temperature lithium batteries, as shown in Fig. 8 [9–11].

Lead-Acid Batteries

The lead-acid (Pb-Acid) battery was invented in 1859. It has been a successful commercial product for over a century. The Pb-Acid battery is mature and has low cost. It has a nominal cell voltage of 2 V, specific energy of 35 Wh/kg, energy density of 90 Wh/L, and specific density of 200 W/kg. It uses metallic lead as the negative electrode and lead dioxide as the positive electrode. The electrolyte is a sulfuric acid solution [16].

On discharge, both lead and lead dioxide are converted into lead sulfate. On charge, the reactions are reversed. The overall electrochemical reactions are described in (1). The electrolyte, sulfuric acid, participates in the electrochemical reactions, and its concentration changes with SoC. Open-circuit voltage of the Pb-Acid battery cell depends only on the acid concentration and is independent of the amount of lead, lead dioxide, or lead sulfate in the cell as long as these components are available. On discharge, the cut-off voltage at moderate rates is 1.75 V and can be as low as 1.0 V at extremely high rates at low temperature. On charge, the charge voltage is regulated below the gassing voltage, about 2.45 V, to avoid evolutions of hydrogen and oxygen gases with the loss of water.

$$\mathrm{Pb} + \mathrm{PbO_2} + 2\mathrm{H_2O} \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} 2\mathrm{PbSO_2} + 2\mathrm{H_2O} \quad (1)$$

Vehicle Energy Storage: Batteries. Table 9 Typical USABC goals for batteries in HEV applications

Parameters	Unit	MHV	MHEV	FHEV	FCEV
Discharge pulse power	kW	6	13–25	40	20
Regenerative pulse power	kW	N.A.	8–20	35	25
Energy capacity	kWh	0.25	0.30	0.50	0.25
Calendar life	Year	15	15	15	15
Cycle life	Cycle	150,000	300,000	300,000	N.A.
Maximum operating voltage	V	48	400	400	440
Operating temperature	°C	−30 to +52	−30 to +52	−30 to +52	−30 to +52

N.A.: not applicable

Vehicle Energy Storage: Batteries. Table 10 Typical USABC goals for batteries in EV applications

Parameters	Unit	EV Commercialization goals	EV Long-term goals
Discharge specific power at 80% DoD for 30 s	W/kg	300	400
Regenerative specific power at 20% DoD for 10 s	W/kg	150	200
Power density	W/L	460	600
Onboard energy capacity	kWh	40	40
Specific energy at C/3 discharge rate	Wh/kg	150	200
Energy density at C/3 discharge rate	Wh/L	230	300
Calendar life	Year	10	10
Cycle life to 80% DoD	Cycle	N.A.	1,000
Operating temperature	°C	−40 to +50	−40 to +85
Selling price	USD/kWh	<150	<100
Normal recharge time	Hour	6	3 to 6
High recharge rate	Hour	0.5 (20–70% SoC)	0.25 (40–80% SoC)

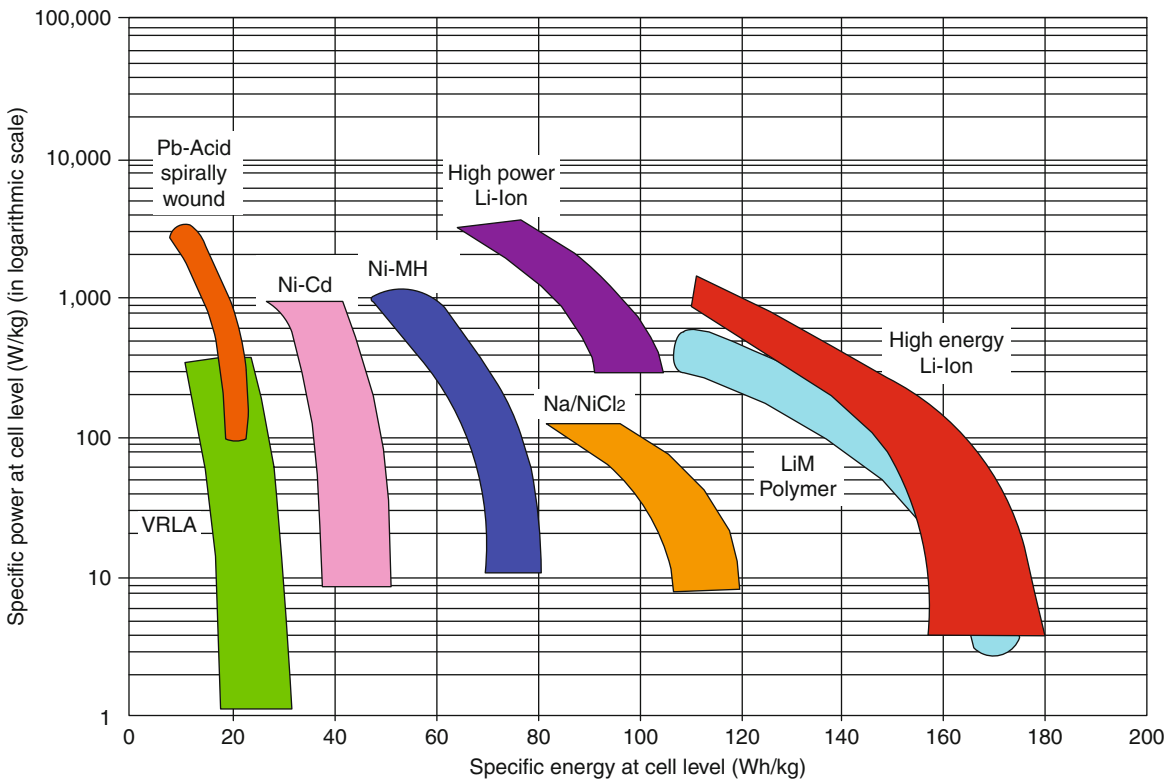
N.A.: not applicable

Valve-Regulated Lead-Acid Battery In the sealed Pb-Acid battery, a special porous separator is employed in the cell such that the evolved oxygen is transferred from the negative electrode to the positive electrode and then combines with hydrogen to form water. Thus, it provides a definite advantage of maintenance-free operation. Moreover, the immobilization of the gelled (Gel) electrolyte or absorbed electrolyte with absorptive glass mat (AGM) separators allows the battery to operate in different orientations without spillage. The sealed Pb-Acid battery is so-called valve-regulated lead-acid (VRLA) battery [16].

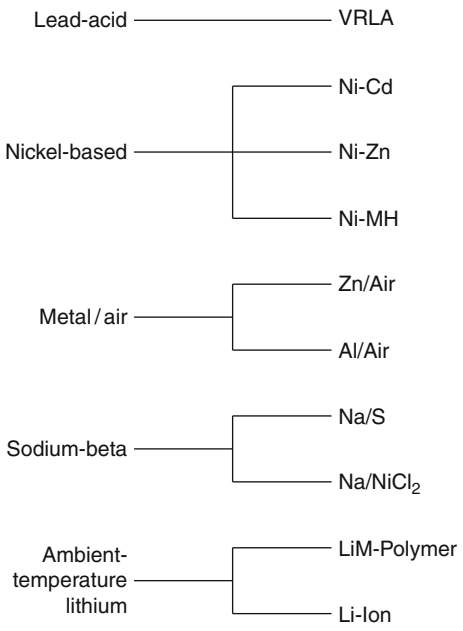
Absorptive Glass Mat The AGM separator in the VRLA batteries serves not only as a permeable electronic insulating diaphragm, resistant to strong acid and oxidation, but also as an acid reservoir for the electrochemical reactions. The separator also plays an active role and has a critical influence on the battery performance and service life. The AGM VRLA battery

Vehicle Energy Storage: Batteries. Table 11 Typical USABC goals for batteries in PHEV applications

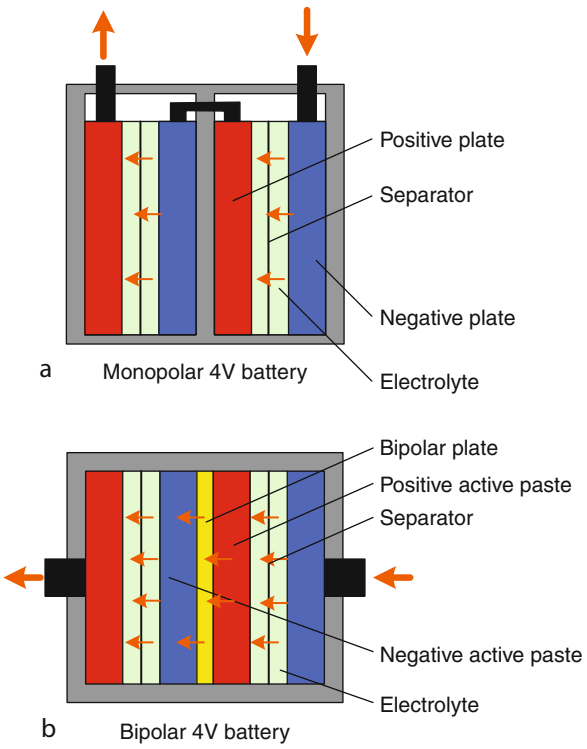
Parameters	Unit	High power to energy ratio battery	High energy to power ratio battery
Reference equivalent electric range	Mile	10	40
Peak pulse discharge power for 2 s	kW	50	46
Peak regenerative power for 10 s	kW	30	25
Available energy for charge-depleting mode at 10 kW discharge rate	kWh	3.40	11.60
Available energy for charge-sustaining mode	kWh	0.50	0.30
Cold cranking power at −30°C	kW	7	7
Calendar life at 35°C	Year	15	15
Maximum system weight	kg	60	120
Maximum system volume	l	40	80
Maximum operating voltage	V	400	400
Operating temperature	°C	−30 to +52	−30 to +52



Vehicle Energy Storage: Batteries. Figure 7
Specific energy and specific power of vehicle batteries



Vehicle Energy Storage: Batteries. Figure 8
Classification of vehicle batteries



Vehicle Energy Storage: Batteries. Figure 9
Conduction paths inside a monopolar battery and inside a bipolar Pb-Acid battery

has penetrated the market of 12 V starting-lighting-ignition (SLI) batteries in premium cars and MHVs. They are attractive for HEBs and low-cost MHVs [17].

Gel Electrolyte In Gel VRLA batteries, the electrolyte is absorbed in a silica gel rather than an AGM. Ionic conductivity of the gelled electrolyte is low such that power density of the Gel VRLA battery is lower than that of the AGM or flooded Pb-Acid batteries. The Gel VRLA battery is not appropriate for starter batteries or power-optimized HEV batteries. Cycling capability of Gel VRLA battery is good but the discharge rate at low temperature is low. Therefore, Gel VRLA batteries are widely used in electric bicycles, BEVs, and in-house transportation systems, where cold cranking is not required but cyclic stress is extreme [18].

Bipolar Cell Stacks Conventional monopolar cell operates at approximately 2 V in its own plastic compartment and the system battery voltage is achieved by connecting a sufficient number of cells in series. In bipolar Pb-Acid batteries, the positive active material of a battery cell is pasted on one side of a conductive plate, so-called bipolar plate, and the negative active material is pasted on the other side. The positive material on one side of a bipolar plate faces the negative material of the neighboring plate with a separator between them. The battery capacity is determined by surface area of the bipolar plate and the paste material utilization [19]. Figure 9 shows the structure and conduction paths in a monopolar battery and in a bipolar Pb-Acid battery.

Bipolar Pb-Acid cells are favorable for high-voltage applications. In the bipolar Pb-Acid battery system, the system voltage increases by 2 V per unit of additional electrode. The bipolar construction shortens the current path between the positive and negative terminals of adjacent cells of the battery. This reduces the battery's internal resistance and creates uniform current distribution such that the paste materials are utilized efficiently and the power intensity is improved.

UltraBattery™ The UltraBattery™ is a hybrid energy storage battery that integrates an asymmetric supercapacitor and a Pb-Acid battery in a single unit

without extra electronic control. The Pb-Acid component comprises one positive plate, lead dioxide (PbO_2), and one negative plate, lead (Pb). The asymmetric supercapacitor consists of one lead dioxide positive electrode and one carbon-based negative electrode. Lead dioxide is the common material of the Pb-Acid cell and the asymmetric supercapacitor such that negative electrodes of the Pb-Acid cell and the asymmetric supercapacitor are connected in parallel and share the same positive electrode in the UltraBattery™ as shown in Fig. 10 [20].

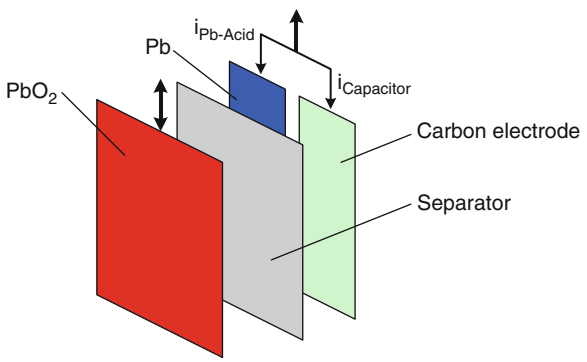
The charge and discharge currents at the negative electrode consists of two components: capacitor current ($i_{Capacitor}$) and Pb-Acid negative electrode current ($i_{Pb\text{-}Acid}$). The capacitor electrode acts as buffer of the Pb-Acid electrode to mitigate peak charge and discharge currents in HEV applications.

The results from tests at the CSIRO laboratory demonstrated that the UltraBattery™ has greater charge and discharge power and significantly long cycle life than that of a traditional VRLA battery. The UltraBattery™ is at the preproduction stage and prototype batteries have been produced at the Furukawa Battery Co., Ltd, Japan for field testing in HEVs.

Vehicle Applications The VRLA battery has maintained its prime position for more than a century. There are a number of advantages contributing to this outstanding position: proven technology and mature manufacturing, low cost, high cell voltage, good high-rate performance that is suitable for vehicle applications, good low-temperature and high-temperature performances, high energy efficiency (75–80%), and availability in a variety of sizes and designs.

The VRLA battery still suffers from some disadvantages and needs continual development. Its specific energy and energy density are relatively low, typically, 35 Wh/kg and 70 Wh/L. Its self-discharge rate is relatively high at about 1% per day at 25°C.

Advanced Pb-Acid batteries with improved performance are being developed for vehicle applications. Improvements of the VRLA battery in specific energy over 40 Wh/kg and energy density over 80 Wh/L with the possibility of rapid recharge have been attained. The bipolar VRLA battery and UltraBattery™ are promising Pb-Acid batteries for vehicle applications.



Vehicle Energy Storage: Batteries. Figure 10
Configuration of an UltraBattery™

Nickel-Based Batteries

There are many kinds of electrochemical batteries using nickel oxyhydroxide as the active material for the positive electrode, including the Ni-Cd, Ni-Zn, and Ni-MH. Among them, the Ni-MH battery has been well accepted for EV and HEV applications because of its proven technology and good performance. The Ni-Zn battery is still under development.

Ni-Cd Battery For more than 90 years, the Ni-Cd battery has been successfully utilized in heavy-duty industrial applications. Due to the resurgence of interest in EVs in the late 1970s and early 1980s, it led to further development of the Ni-Cd battery for EV applications. The Ni-Cd battery possesses the nominal parameters of 1.3 V, 56 Wh/kg, 110 Wh/L, and 225 W/kg. Its active materials are metallic cadmium for the negative electrode and nickel oxyhydroxide for the positive electrode. The alkaline electrolyte is an aqueous potassium hydroxide solution [21].

The electrochemical reactions of discharge and charge are described in (2). On discharge, metallic cadmium is oxidized to form cadmium hydroxide and nickel oxyhydroxide is reduced to nickel hydroxide with consumption of water. On charge, the reverse reactions occur. In contrast to the sulfuric acid electrolyte used in the Pb-Acid battery, the potassium hydroxide electrolyte in the Ni-Cd battery is not significantly changed in density or composition during discharge and charge.

$$2\mathrm{NiOOH} + \mathrm{Cd} + 2\mathrm{H_2O} \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} 2\mathrm{Ni(OH)_2} + \mathrm{Cd(OH)_2} \quad (2)$$

The Ni-Cd battery has gained enormous technical importance because of the advantages of high specific power (over 220 W/kg), long cycle life (up to 2,000 cycles), highly tolerant of electrical and mechanical abuse, flat voltage profile over a wide range of discharge currents, rapid recharge capability (about 40–80% in 18 min), wide operating temperature range (from −40°C to 85°C), low self-discharge rate (less than 0.5% per day), excellent long-term storage due to negligible corrosion, and available in a variety of sizes and designs. The Ni-Cd battery has some disadvantages which offset its wide acceptance for vehicle applications, namely, low cell voltage, memory effect, and the carcinogenicity and environmental hazard of cadmium.

The Ni-Cd battery can be divided into vented and sealed types. The vented sintered-plate type has higher specific energy but is more expensive. It is characterized by its flat discharge profile and superior high-rate and low-temperature performance. Similar to the sealed Pb-Acid battery, the sealed Ni-Cd battery incorporates a specific cell design feature to prevent build-up of pressure in the cell caused by gassing during overcharge. As a result, the battery can be sealed and requires no maintenance other than recharging.

Major manufacturers of the Ni-Cd battery for vehicle applications are SAFT and VARTA. EVs powered by the Ni-Cd battery included the Chrysler TE Van, Citroën AX, Mazda Roadster, Mitsubishi EV, Peugeot 106, Renault Clio, and HKU U2001.

Ni-Zn Battery Starting from the 1930s, the Ni-Zn battery has been studied for vehicle applications. The Ni-Zn battery has high specific energy and low material cost; however, it has not achieved any commercial importance because of the short life in the zinc electrode [22].

The Ni-Zn battery nominally operates at 1.6 V and has energy and power densities of 60 Wh/kg, 120 Wh/L, and 300 W/kg. It uses zinc as the negative electrode and nickel oxyhydroxide as the positive electrode. The electrolyte is an alkaline potassium hydroxide solution. The discharge and charge reactions are described in (3).

$$\mathrm{Zn} + 2\mathrm{NiOOH} + 2\mathrm{H_2O} \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} \mathrm{Zn(OH)_2} + 2\mathrm{Ni(OH)_2} \quad (3)$$

On discharge, metallic zinc in the negative electrode is oxidized to form zinc hydroxide and nickel oxyhydroxide in the positive electrode is reduced to nickel hydroxide. On charge, the reactions are reversed.

Among the nickel-based batteries, the Ni-Zn battery has the advantages of higher specific energy and specific power than the Ni-Cd battery's, high cell voltage (the highest of the nickel-based family), nontoxicity (more environmental friendliness than the Ni-Cd), tolerance of overcharge and overdischarge, capability of high discharge and recharge rates, and wide operating temperature (from −20°C to 60°C). However, the major and serious drawback of the Ni-Zn battery is its short cycle life (about 300 cycles). It is mainly due to the partial solubility of zinc species in the electrolyte.

Ni-MH Battery The Ni-MH battery has been on the market since 1992. Its characteristics are similar to the Ni-Cd battery. The principal difference between them is the use of hydrogen, absorbed in a metal hydride, for the active negative electrode material in the Ni-MH battery [23].

Active materials of Ni-MH batteries are hydrogen in the form of metal hydride for the negative electrode and nickel oxyhydroxide for the positive electrode. The metal hydride undergoes reversible hydrogen desorbing-absorbing reactions when the battery is discharged and recharged. An aqueous solution of potassium hydroxide is the major component of the electrolyte. The overall electrochemical reactions are described in (4). When the battery is discharging, metal hydride in the negative electrode is oxidized to form metal alloy and nickel oxyhydroxide in the positive electrode is reduced to nickel hydroxide. During charge, the reverse reactions occur.

$$\mathrm{MH} + \mathrm{NiOH} \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} \mathrm{M} + \mathrm{Ni(OH)_2} \qquad (4)$$

The hydrogen storage metal alloy is a key component of the Ni-MH battery. It is well formulated to maintain stable over a large number of cycles. The rare-earth alloys based around lanthanum nickel, known as the AB5, and the alloys consisting of titanium and zirconium, known as the AB2, are the two major metal alloys used in Ni-MH batteries. Although the AB2 alloys typically have higher capacity than that of the AB5 alloys, the AB5 alloy is widely used because of its superiority in charge retention and stability.

The Ni-MH battery has a nominal voltage of 1.32 V and attains specific energy of 65–110 Wh/kg for EV applications and 45–60 Wh/kg for HEV applications. It operates in a temperature from −20°C to +45°C. A number of battery manufacturers, such as GM Ovonic, GP, GS, Panasonic, SAFT, VARTA, and YUASA, have actively engaged in the development of this battery for HEVs.

Metal/Air Batteries

The rechargeable metal/air batteries include the electrically or mechanically rechargeable zinc/air (Zn/Air) battery and the mechanically rechargeable aluminum/air (Al/Air) battery. These metal/air batteries have very high specific energy and energy density (as high as 600 Wh/kg and 400 Wh/L for Al/Air), low cost, and are environment friendly. In addition, those mechanically rechargeable batteries have two distinct advantages which are very essential for EV applications, namely, fast and convenient refueling (comparable to petrol refueling in a few minutes) and centralized recharging and recycling (most efficient and environmentally sound use of electricity). The disadvantages associated with rechargeable metal/air batteries are low specific power (at most 105 W/kg for Zn/Air), narrow operating temperature window, carbonation of alkaline electrolyte due to carbon dioxide in air, and evolution of hydrogen gas from corrosion in electrolyte.

Zn/Air Battery The Zn/Air battery has been developed as an electrically and mechanically rechargeable battery. Although both of them have been applied to EV applications, the mechanically rechargeable battery is more favorable [24].

The electrically rechargeable Zn/Air battery nominally operates at 1.2 V and has the specific energy of 180 Wh/kg, energy density of 160 Wh/L, and specific power of 95 W/kg. The negative electrode consists of zinc particles and the positive electrode is a bifunctional air electrode. The electrolyte is potassium hydroxide. The simplified electrochemical reactions are described in (5).

$$2\mathrm{Zn} + \mathrm{O_2} \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} 2\mathrm{ZnO} \qquad (5)$$

On discharge, zinc is first oxidized to potassium zincate dissolved in the electrolyte and then to a precipitation of zinc oxide. On charge, the reactions are reversed. However, the zinc electrode generally suffers from the problem of shape change during cycling.

The mechanically rechargeable Zn/Air battery avoids the use of bidirectional air electrode and the shape change problem. Hence, it can offer a higher specific energy of 230 Wh/kg and a higher specific power of 105 W/kg. The depleted zinc negative electrode cassettes can be replaced robotically by a mechanically refueling system at a fleet servicing point or at a public service station. The discharged fuel is then electrochemically recharged at central facilities. There are four steps in a recharging process. Firstly, the discharged cassettes are mechanically taken apart and the zinc oxide discharge product is removed. Secondly, zinc oxide is dissolved in a potassium hydroxide solution to form a zincate solution. Thirdly, the zincate solution is electrolyzed in an electrowinning bath. Finally, the electrowon zinc is compacted onto the negative electrode cassettes.

A mechanically rechargeable Zn/Air battery was developed for field test. A 160-kWh Zn/Air battery was installed and tested in a Mercedes-Benz 180E van in 1994. The driving range at a constant speed of 64 km/h was 689 km.

Al/Air Battery The Al/Air battery has a nominal voltage of 1.4 V. The negative electrode is aluminum metal and the positive electrode is only a simple

unifunctional air electrode for discharge. The electrolyte can be either saline solution or alkaline potassium hydroxide solution. The discharge reaction of this mechanically rechargeable battery is described in (6).

$$4Al + 3O_2 + 6H_2O \xrightarrow{\text{discharge}} 4Al(OH)_3 \qquad (6)$$

The Al/Air battery with a saline electrolyte is attractive only for low-power applications. On the other hand, the alkaline Al/Air battery offers high specific energy and energy density of 250 Wh/kg and 200 Wh/L and is suitable for high power applications. Nevertheless, the corresponding specific power is as low as 7 W/kg. Because of its exceptionally low specific power, the Al/Air battery is seldom used as the sole energy source for EVs and it is commonly used in conjunction with other batteries in a battery hybrid system.

Sodium-Beta Batteries

The sodium-beta battery refers to the Na/S and $Na/NiCl_2$ batteries, which have liquid sodium as one reactant and beta-alumina ceramic as the electrolyte.

Na/S Battery The Na/S battery operates at 300–350°C with a nominal cell voltage of 2 V, specific energy of 170 Wh/kg, energy density of 250 Wh/L, and specific power of 390 W/kg. The active materials are molten sodium for the negative electrode and molten sulfur/sodium polysulfides for the positive electrode. The beta-alumina ceramic electrolyte functions as a sodium ion-conducting solid medium and works as a separator for the molten electrodes to prevent any direct self-discharge [25]. The electrochemical reactions of the Na/S battery are described in (7).

$$2Na + xS \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} Na_2S_x, \quad x = 2.7 - 5 \qquad (7)$$

On discharge, sodium is oxidized to form sodium ions, which migrate through the electrolyte and combine with the sulfur that is being reduced in the positive electrode to form sodium pentasulfide. Then, the sodium pentasulfide is progressively converted into polysulfides with higher sulfur compositions (Na_2S_x) where x is from 2.7 to 5. On charge, these reactions are reversed.

Barriers to commercialization of the Na/S battery are safety issues (high reactivity and corrosiveness of molten active materials), inadequate freeze-thaw durability (weak ceramic electrolyte subjected to mechanical stress), and need of thermal management (additional energy and thermal insulation).

$Na/NiCl_2$ Battery In $Na/NiCl_2$ battery, the active materials are molten sodium for the negative electrode and solid nickel chloride for the positive electrode. In addition to the beta-alumina ceramic electrolyte as used in the Na/S, there is a secondary electrolyte, namely, sodium-aluminum chloride, in the positive electrode chamber. The secondary electrolyte conducts sodium ions from the primary beta-alumina electrolyte to the solid nickel chloride positive electrode [26]. The corresponding electrochemical reactions are described in (8)

$$2Na + NiCl_2 \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} Ni + 2NaCl \qquad (8)$$

On discharge, the solid nickel chloride is converted into nickel metal and sodium chloride crystal. On charge, these reactions are reversed. The $Na/NiCl_2$ battery operates at 155–350°C with a nominal cell voltage of 2.58 V. In a battery system, the system specific energy and specific power can be 86–120 Wh/kg and 150–300 W/kg. Comparing with the Na/S battery, the $Na/NiCl_2$ battery has higher open circuit cell voltage, wider operating temperature, safer products of reaction (less corrosive than molten Na_2S_x), and better freeze-thaw durability (smaller temperature difference).

The AEG ZEBRA (Zero Emission Battery Research Activity) has been the major developer of the $Na/NiCl_2$ battery. The ZEBRA battery, namely, Z12, offered a specific energy of 103 Wh/kg and a specific power of 180 W/kg.

Ambient Temperature Lithium Batteries

There are a number of approaches being taken in the design of rechargeable ambient-temperature lithium

batteries. One approach is to use metallic lithium for the negative electrode and a solid inorganic intercalation material for the positive electrode. The electrolyte can be a solid polymer, leading to name as the lithium metal polymer (LiM-Polymer) battery. Another approach is the use of a lithiated carbon material as the negative electrode such that lithium ions move forth and back between the positive and negative electrodes during cycling. The "rocking-chair" movements of Li-Ions lead to the name lithium-ion (Li-Ion) battery [1, 27].

Lithium Metal Polymer Batteries The LiM-Polymer battery uses lithium metal and a transition metal intercalation oxide (M_yO_z) for the negative and positive electrodes, respectively. A thin solid polymer electrolyte (SPE) is used, which offers the merits of improved safety and flexibility in design [28]. The general electrochemical reactions are described in (9).

$$x\mathrm{Li} + \mathrm{M}_y\mathrm{O}_z \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} \mathrm{Li}_x\mathrm{M}_y\mathrm{O}_z \tag{9}$$

On discharge, lithium ions formed at the negative electrode migrate through the SPE and are inserted into the crystal structure at the positive electrode. On charge, the process is reversed. By using a lithium foil negative electrode and vanadium oxide (V_6O_{13}) positive electrode, the Li/SPE/V_6O_{13} cell is a typical LiM-Polymer battery. It operates at a nominal voltage of 3 V and has the specific energy of 155 Wh/kg, energy density of 220 Wh/L, and specific power of 315 W/kg. The advantages are high cell voltage (3 V), very high specific energy and energy density (155 Wh/kg and 220 Wh/L), very low self-discharge rate (about 0.5% per month), and capability of fabrication in a variety of shapes and sizes. However, its low-temperature performance is weak [1].

Lithium Ion Batteries Since the commercialization of the Li-Ion battery by Sony Energytec in 1990, the Li-Ion battery has been considered to be the most promising rechargeable battery of the future. The Li-Ion battery has already gained acceptance for HEV applications. The specific energy of Li-Ion battery was 98 Wh/kg in 1990 and increased to 195 Wh/kg in 2008 [27].

The Li-Ion battery consists of two electrodes, a porous separator impregnated with electrolyte, and two current collectors. Lithium cobalt oxide ($LiCoO_2$) typically serves as an active electrode material for the positive electrode. The negative electrode is usually made of lithiated carbon or graphite (LiC_6). Electrodes are electrically isolated by the separator, where the space between them is filled by electrolyte. Copper foil is used for the negative current collector and aluminum for the positive current collector.

The Li-Ion battery employs insertion reactions for both positive and negative electrodes. The Li-Ions are inserted into the negative electrode when the battery is fully charged and they move between positive and negative electrodes during cycling. Thus, the Li-Ion battery is also called "rocking-chair battery" or "shuttle-cock battery." The main reactions at the positive electrode are described in (10).

$$\mathrm{Li}_{1-x}\mathrm{CoO}_2 + x\mathrm{Li}^+ + xe^- \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} \mathrm{LiCoO}_2 \quad 0 \le x \le 0.5 \tag{10}$$

The Li-Ion is extracted from the positive electrode during charging and inserted into the electrode during discharging, where x represents the mol fraction of Li^+ ions inside the positive electrode. For a proper reversible functioning of a Li-Ion battery, not all Li^+ ions can be extracted from the electrode. The corresponding reactions at the negative electrode are described in (11).

$$\mathrm{Li}_z\mathrm{C}_6 \underset{\text{charge}}{\overset{\text{discharge}}{\rightleftharpoons}} \mathrm{C}_6 + z\mathrm{Li}^+ + ze^- \quad 0 \le z \le 1 \tag{11}$$

Where z describes the mol fraction of Li^+ ions inside the negative electrode.

The electrolyte is based on a dissociated lithium-containing salt, such as lithium hexafluorophosphate ($LiPF_6$) or lithium perchlorate ($LiClO_4$). Mixtures of ethylene carbonate (EC), diethyl carbonate (DEC), and dimethyl carbonate (DMC) are used as nonaqueous solvents. The ions in the electrolyte are transported by both diffusion and migration, which is induced by the

electric field between the electrodes. The overall reactions are described in (12)

$$\mathrm{Li}_{1-x}\mathrm{CoO}_2 + \mathrm{Li}_{z+x}\mathrm{C}_6 \underset{\text{charge}}{\overset{\text{discharge}}{\longleftrightarrow}} \mathrm{LiCoO}_2 + \mathrm{Li}_z\mathrm{C}_6 \tag{12}$$

Development of Positive Active Materials The $LiCoO_2$ has been used as positive active material since 1990. Extensive works have been done for new materials with higher energy density and lower cost. Electrode with nickel compounds shows larger capacity than that of $LiCoO_2$. The lithium nickel oxide ($LiNiO_2$) is a potential material for the positive electrode of the Li-Ion battery. The $LiNiO_2$ is a layered oxide with lithium, oxygen, and nickel layers, and it has the same crystalline structure of $LiCoO_2$. The $LiNiO_2$ is not commercialized because the $LiNiO_2$ electrode showed a rapid capacity decay with cycling, due to the movement of nickel to the lithium layer during cycling. In addition, the thermal stability of $LiNiO_2$ at charged state is low such that it goes to thermal runaway easily. Partial substitutions of nickel with Co, Al, and Mn have been studied to enhance thermal stability.

The manganese-based compounds, namely, $LiMnO_2$ and $LiMn_2O_4$ are potential materials for the positive electrode. The three-dimensional structure and spinel $LiMn_2O_4$ is more stable than the $LiCoO_2$ at fully charged state and the cost is lower. The Li-Ion battery with spinel $LiMn_2O_4$ was commercialized in 1996. However, the battery capacity of $LiMn_2O_4$ is lower than that of $LiCoO_2$ and the capacity decays rapidly. Partial substitutions of $LiMn_2O_4$ with Co, Mg, Cr, Ni, Fe, Ti, and Zn improve cycle life of spinel $LiMn_2O_4$.

The positive electrode with polyanions (SO_4^{2-} and PO_4^{2-}) are promising in reducing cost and enhancing thermal stability. The lithium iron phosphate ($LiFePO_4$) is electrochemically active and shows a flat discharge voltage at about 3.5 V. The $LiFePO_4$ electrode is thermally stable, because it does not release oxygen at fully charged state at an elevated temperature. The electric conductivity is enhanced by coating nanosized $LiFePO_4$ particles with ultrathin carbon layer. Moreover, lithium manganese phosphate ($LiMnPO_4$) and lithium cobalt phosphate ($LiCoPO_4$) are under development for enhancements of battery voltage and capacity [29–34].

Development of Negative Active Materials The capacity of negative electrode has increased by replacing coke (non-graphitic carbon) with graphite, which has larger capacity and flat discharge profile. The capacity of graphic carbon used in Li-Ion batteries is close to the theoretical limit. Commercially available graphic carbons are natural graphite, synthetic graphite, and mesocarbon microbead (MCMB). Intermetallic components and lithium titanium oxide are potential materials for negative electrodes of Li-Ion batteries [35].

Lithium alloys have been developed for negative electrode since 1970s. Cycle life of lithium alloy was short due to alloy pulverization caused by large volume change during cycling. Intermetallic components, such as Cu − Sn, Cu − Sb, and In − Sb, have been investigated aiming at suppressing volume change in lithium alloys during cycling. In 2005, Sony demonstrated a lithium alloy, composed of Sn, Co, and C, that showed 50% increase in volumetric capacity, comparing with a conventional graphite electrode.

The lithium titanate spinel has been investigated as a negative electrode material since the commercialization of the Li-Ion battery, which uses $LiCoO_2$ cathode and carbon anode. The lithium titanate materials, particularly the $Li_4Ti_5O_{12}$, have demonstrated significant improvements in charge and discharge rates in laboratory and in commercial batteries [36].

Vehicle Applications The Li-Ion battery can be made from different advanced positive electrode and negative electrode materials. The mature positive electrode materials are $LiCoO_2$, $LiMn_2O_4$, $LiFePO_4$, lithium nickel manganese cobalt (NMC) oxide ($LiNiMnCoO_2$), and lithium nickel cobalt aluminum (NCA) oxide ($LiNiCoAlO_2$). The mature negative electrode materials are graphite and titanate. Table 12 shows the potential Li-Ion batteries for vehicle applications. Each Li-Ion battery in Table 12 can only overcome one of the barriers to EV commercialization. However, these high-power and thermally stable Li-Ion batteries are promising for HEV applications.

Future Directions

The battery is the most significant factor of commercialization of EVs and HEVs. Developments of batteries for EV and HEV applications are continued and

Vehicle Energy Storage: Batteries. Table 12 Advanced Li-Ion batteries

Positive electrode	Negative electrode	Manufacturers	Key feature
$LiCoO_2$	Graphite	Sony	Mature
$LiMn_2O_4$	Graphite	NEC, GS, Yuasa, LG	High power
NCA/NMC	Graphite	SAFT, Samsung, Sanyo, Evonik	High energy
$LiFePO_4$	Graphite	A123, Valence Tech, BYD	Highly stable
$LiMn_2O_4$	Titanate	Toshiba, Enerdel	High discharge rate

Vehicle Energy Storage: Batteries. Table 13 Key features of promising batteries

Type	Pb-Acid	Nickel-based		Lithium	
Feature	VRLA	Ni-Cd	Ni-MH	Li-Ion	Li-Titanate
Specific energy (Wh/kg)	30–40	40–60	60–70	160	70–90
Cycle life at 100% DoD (cycle)	50–80	300–600	300–500	500–750	25,000
Safety	Fire hazard	Moderate	Fire hazard	Fire hazard	Safest
Charge time (0–90% SoC) (h)	8	2	2	2	0.1
Operating temperature (°C)	−10 to 60	0–50	0–40	0–40	−40 to 70
Environmental impact	Toxic	Toxic	Low	Minimal	Minimal
Memory effect	Very low	High	Moderate	None	None
Power delivery	Good	Moderate	Moderate	Moderate	High
Manufacturability	Easy	Adequate	Adequate	Easy	Easy
Maintenance	Moderate	Moderate	Moderate	Moderate	Moderate
Market position	High volume	Sliding	Modest	Good	Rising
Cost	Low	Tied to Ni	Tied to Ni	Moderate	Moderate

accelerated. The key requirements for vehicular applications are safety, high specific energy, high specific power, short recharge time, long life cycle, and low cost.

The mature and promising batteries for EVs and HEVs are VRLA, Ni-Cd, Ni-MH, and Li-Ion batteries. The specific power and specific energy of the batteries are plotted in Fig. 3 and their features are compared in Table 13. The VRLA battery is popular for MHVs and low-cost EVs due to its maturity and cost-effectiveness. Features of the Ni-MH battery are superior to those of the Ni-Cd battery, so the Ni-Cd battery is being superseded by the Ni-MH battery in the market for MHEVs, FHEVs, and PHEVs.

The Li-Ion batteries are the promising batteries in the future. The advanced Li-Ions batteries have demonstrated the potentials of improvements in specific power, specific energy, charge rate, and safety.

Bibliography

Primary Literature

1. Chan CC, Chau KT (2001) Modern electric vehicle technology. Oxford University Press, Oxford
2. Kurzweil P (2009) Secondary batteries. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 565–578
3. Chau KT, Wong YS (2002) Overview of power management in hybrid electric vehicles. J Energy Conver Manage 43:953–1968
4. Chan CC, Wong YS, Bouscayrol A, Chen K (2009) Powering sustainable mobility: roadmaps of electric, hybrid, and fuel cell vehicles. Proc IEEE 97:603–607

5. Chan CC (2007) The state of the art of electric, hybrid, and fuel cell vehicles. Proc IEEE 95:704–718
6. Chan CC, Wong YS (2004) The state of the art of electric vehicles technology. IPEMC 2004 1:46–57
7. Chau KT, Chan CC (2007) Emerging energy-efficient technologies for hybrid electric vehicles. Proc IEEE 95:821–835
8. Chan CC, Bouscayrol A, Chen K (2010) Electric, hybrid, and fuel-cell vehicles: architectures and modeling. IEEE Trans Veh Technol 59:589–598
9. Köhler U (2009) Hybrid electric vehicles: batteries. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 269–285
10. Chau KT, Wong YS (2001) Hybridization of energy sources in electric vehicles. J Energy Conver Manage 42:1059–1069
11. Gutmann G (2009) Electric vehicle: batteries. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 219–235
12. USCAR (2006) Energy storage system goals: USABC goals for advanced batteries for EVs. USCAR. http://www.uscar.org
13. USCAR (2006) Energy storage system goals: power assist HEV battery goals. USCAR. http://www.uscar.org
14. USCAR (2006) Energy storage system goals: 42 V battery goals. USCAR. http://www.uscar.org
15. USCAR (2008) USABC requirements of end of life energy storage systems for PHEVs. USCAR. http://www.uscar.org
16. Rand DAJ, Moseley PT (2009) Lead-acid system overview. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 550–575
17. Weighall MJ (2009) Valve-regulated batteries: absorptive glass mat. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 715–726
18. Kramm F, Niepraschk H (2009) Valve-regulated batteries: gel. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 727–734
19. Loyns AC (2009) Bipolar batteries. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 750–754
20. Lam LT, Furukawa J (2009) Supercap hybrid (UltraBattery™). In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 755–763
21. Shukla AK, Venugopalan S, Hariprakash B (2009) Nickel–cadmium: overview. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 452–458
22. Cairns EJ (2009) Nickel–Zinc. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 528–533
23. Hariprakash B, Shukla AK, Venugoplan S (2009) Nickel–metal hydride: overview. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 494–501
24. Haas O, Van Wesemael J (2009) Zinc–air: electrical recharge. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 384–392
25. Holze R (2009) Sodium–sulfur. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 302–311
26. Sudworth JL, Galloway RC (2009) Sodium–nickel chloride. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 312–323
27. Yamaki J (2009) Lithium rechargeable systems – lithium-ion overview. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 183–191
28. Kobayashi Y, Seki S, Terada N (2009) Lithium-ion polymer batteries. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 375–382
29. Goodenough JB (2009) Positive electrode: layered metal oxides. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 243–248
30. Arai H, Hayashi M (2009) Positive electrode: lithium cobalt oxide. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 258–263
31. Zaghib K, Mauger A, Gendron F, Julien CM, Goodenough JB (2009) Positive electrode: lithium iron phosphate. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 264–296
32. Kanno R (2009) Positive electrode: lithium nickel oxide. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 297–306
33. Wohlfahrt-Mehrens M (2009) Positive electrode: manganese spinel oxides. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 318–327
34. Liu DW, Cao GZ, Wang Y (2009) Positive electrode: nanostructured transition metal oxides. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 337–355
35. Inaba M (2009) Negative electrodes: graphite. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 198–208
36. Ariyoshi K, Ohzuku T (2009) Negative electrode: spinel-type titanium oxides. In: Encyclopedia of electrochemical power sources. Elsevier, Amsterdam, pp 209–213

Books and Reviews

Encyclopedia of electrochemical power sources. Elsevier

Vehicle Traction Motors

C. C. Chan[1], Ming Cheng[2]
[1]Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China
[2]School of Electrical Engineering, Southeast University, Nanjing, China

Article Outline

Glossary

AC motor An electric motor driven by an alternating current. There are two types of AC motors, depending on the type of rotor used. The first is the synchronous motor, which rotates exactly at the supply frequency or a submultiple of the supply frequency. The magnetic field on the rotor is either generated by current delivered through slip rings or by a permanent magnet. The second is the induction motor, which runs slightly slower than the supply frequency. The magnetic field on the rotor of this motor is created by an induced current.

Armature winding The conducting coils that are wound around the armature in which voltage is induced, causing it to rotate within a magnetic field.

Brushless DC motor Also called electronically commutated motors. Synchronous motors powered by direct current supply and having electronic commutation system, rather than mechanical commutators and brushes. The current-to-torque and voltage-to-speed relationships are linear.

CVT Continuous variable transmission is a transmission which can change steplessly through an infinite number of effective gear ratios between maximum and minimum values. This contrasts with other mechanical transmissions that only allow a few different distinct gear ratios to be selected. The flexibility of a CVT allows the driving shaft to maintain a constant angular velocity over a range of output velocities.

DC motor An electric motor that runs on direct current (DC) supply.

DTC Direct torque control is a method used in variable frequency drives to control the torque of three-phase AC motors based on stator flux control in the stator fixed frame using direct control of the inverter switching. It involves estimating the motor's magnetic flux and torque based on the measured voltage and current of the motor.

emf Electromotive force is the force that pushes electrons through a conductor.

Field winding The electric circuit is usually a number of coils wound on individual poles and connected in series, which produces the magnetic field in a motor or generator.

FOC Field-oriented control, also called vector control, is a method used in variable frequency drives to control the torque (and thus finally the speed) of three-phase AC motors by controlling two orthogonal current vectors.

Generator A machine that converts mechanical energy into electrical energy by magnetic induction.

ISG Integrated starter/generator, an advanced electric machine controlled by electronics and is designed for integration with internal combustion engines. It replaces the conventional starter motor and alternator, which are the two indispensable electric units for almost every engine.

mmf Magnetomotive force, also known as magnetic potential, is the property of certain substances or phenomena that give rise to magnetic fields. Magnetomotive force is analogous to electromotive force or voltage in electric field.

Motor A machine that converts one form of energy, such as electricity, into mechanical energy or motion.

Definition of the Subject

The traction motor of EVs is responsible for converting electrical energy to mechanical energy in such a way that the vehicle is propelled to overcome aerodynamic drag, rolling resistance drag, and kinetic resistance.

Some engineers and even researchers may consider traction motors kindred or similar to industrial motors. However, traction motors usually require frequent start/stop, high rate of acceleration/deceleration, high-torque low-speed hill climbing, low-torque high-speed cruising, and very wide speed range of operation, whereas industrial motors are generally optimized at rated conditions. Thus, traction motors are so unique that they are deserved to form an individual class. Hence, the general requirements of traction motor are significantly different from those of industrial motors. Their major differences in load

requirement, performance specification, and operating environment are as follows:

- Traction motors need to offer the maximum torque that is four to five times of the rated torque for temporary acceleration and hill climbing, while industrial motors generally offer the maximum torque that is twice of the rated torque for overload operation.
- Traction motors need to achieve four to five times the base speed for highway cruising, while industrial motors generally achieve up to twice the base speed for constant-power operation.
- Traction motors should be designed according to the vehicle driving profiles and drivers' habits, while industrial motors are usually based on a typical working mode.
- Traction motors demand both high power density and good efficiency map (high efficiency over wide speed and torque ranges) for the reduction of total vehicle weight and the extension of driving range, while industrial motors generally need a compromise among power density, efficiency, and cost with the efficiency optimized at a rated operating point.
- Traction motors desire high controllability, high steady-state accuracy, and good dynamic performance for multiple-motor coordination, while only special-purpose industrial motors desire such performance.
- Traction motors need to be installed in mobile vehicles with harsh operating conditions such as high temperature, bad weather, and frequent vibration, while industrial motors are generally located in fixed places.

Introduction

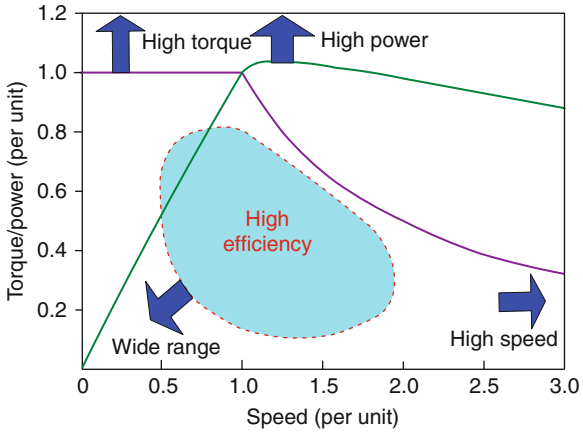
An electric motor drive is the heart of electric vehicles (EVs). Its job is to interface energy source (such as batteries) with vehicle wheels, transferring energy in either direction as required, with high efficiency, under control of the driver at all times. Hence, the electric motor drives are the core technology for electric, hybrid, and fuel cell vehicles. The major requirements of the traction motor drive are the following: [1–3]:

1. High torque density and power density
2. Very wide speed range, including constant-torque and constant-power regions
3. High efficiency over wide torque and speed ranges
4. High torque for low-speed starting and climbing, and high power for high-speed cruising
5. Fast torque response
6. High intermittent overload capability for overtaking
7. High reliability and robustness for vehicular environment
8. Low acoustic noise
9. Reasonable cost

Typical torque/power-speed characteristics required for traction motor drives are illustrated in Fig. 1.

To satisfy these special requirements, the power rating and torque-speed requirements of the motor drive should be determined on the basis of driving cycles and system-level consideration. New motor design technologies and control strategies are being pursued to extend the speed range, to optimize the system efficiency, and to enlarge the high-efficiency region. Newly developed electronic products are also adopted to improve the system performance and to reduce the total cost.

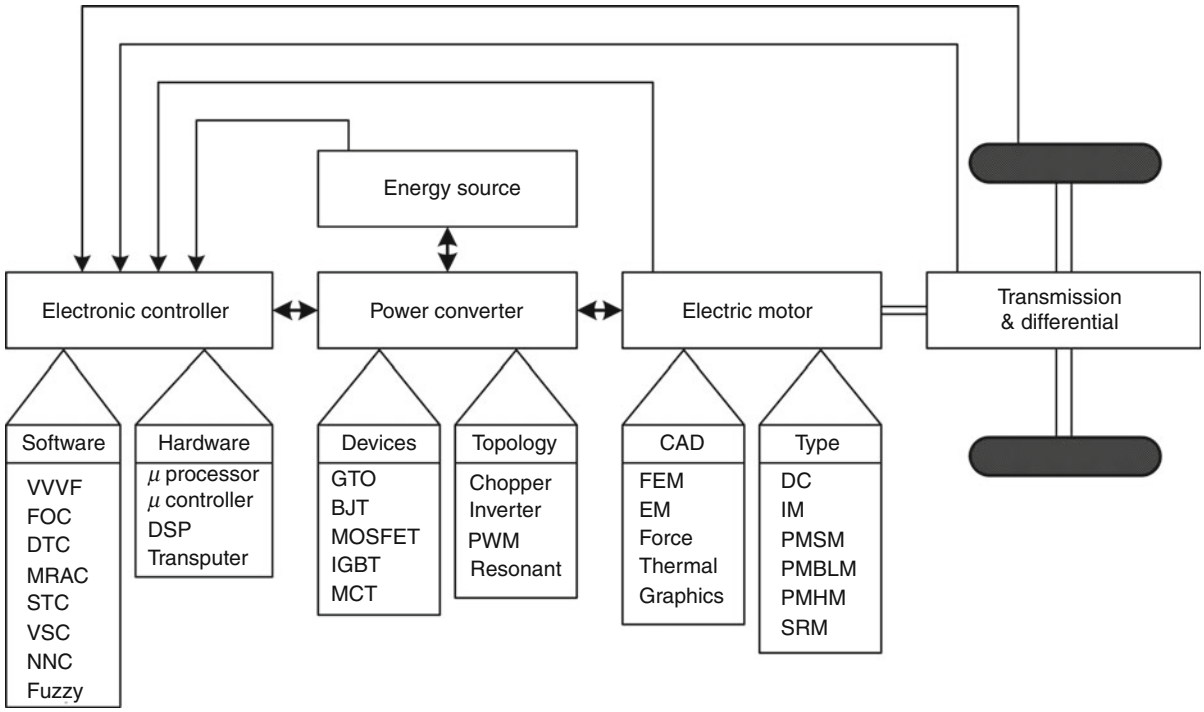
From the functional point of view, a traction motor drive system can be divided into two parts – electrical and mechanical. The electrical part consists of the subsystems of motor, power converter, and electronic controller, whereas the mechanical part includes the subsystems of mechanical transmission (optional) and vehicle wheels. The boundary between the



Vehicle Traction Motors. Figure 1
Torque/power requirements for traction motors

electrical and mechanical parts is the air-gap of the motor, where electromechanical energy conversion is taken place. The electronic controller can be further divided into three functional units – sensor, interface circuitry, and processor. The sensor is used to translate the measurable quantities, such as current, voltage, temperature, speed, torque, and flux, into electronic signals. Through the interface circuitry, these signals are conditioned to the appropriate level before being fed into the processor. The processor output signals are usually amplified via the interface circuitry to drive power semiconductor devices of the power converter. The converter acts as a power conditioner that regulates the power flow between the energy source and the electric motor for motoring and regeneration. Finally, the motor interfaces with the vehicle wheels via the mechanical transmission. This transmission is optional because the electric motor can directly drive the wheel as in the case of in-wheel drives. The functional block diagram of a motor drive for EVs is shown in Fig. 2.

Based on the technological growth of electric motors, power electronics, microelectronics, and control strategies, more and more kinds of motor drives become available for EVs. DC motor drives have been traditionally used for EV propulsion because of their ability to achieve high torque at low speeds and easy control. However, the DC motor needs careful maintenance due to its commutator and brushes. Recent technological developments have enabled a number of advanced motor drives to offer definite advantages over those DC motor drives, namely, high efficiency, high power density, efficient regenerative braking, robust, reliable, and maintenance free. Among them, the vector controlled induction motor drive is most popular and mature, though it may suffer from low efficiency at light-load ranges. On the other hand, permanent magnet (PM) brushless motors possess the highest efficiency and power density over the others, but may suffer from a difficulty in flux weakening control for the constant-power high-speed region. The PM hybrid motor is a special type of PM brushless motors. In this motor, an auxiliary DC field winding is so incorporated that the air-gap flux is a resultant of the PM flux and field-winding flux.



Vehicle Traction Motors. Figure 2
Functional block diagram of an EV drive system

By adjusting the field-winding excitation current, the air-gap flux can be varied flexibly, hence offering optimal efficiency over a wide speed range. Switched reluctance (SR) motors offer promising features for EV applications due to their simplicity and reliability in both motor construction and power converter configuration, wide speed range, favorable thermal distribution, and efficient regenerative braking. However, they may suffer from torque ripples and acoustic noise problems.

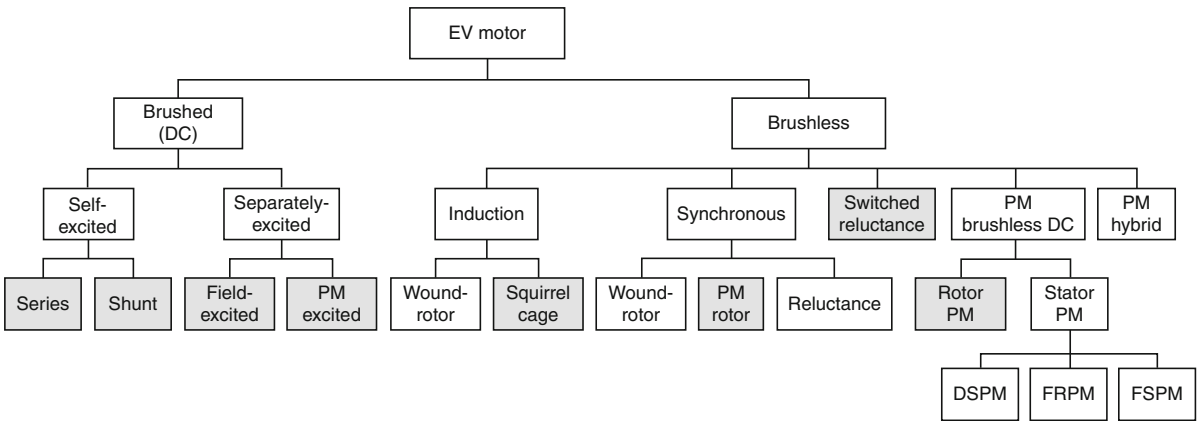
The choice of traction motor drive for EVs mainly depends on three factors – driver expectation, vehicle constraint, and energy source. The driver expectation is defined by a driving profile which includes the acceleration, maximum speed, climbing capability, braking, and range. The vehicle constraint depends on the vehicle type, vehicle weight, and payload. The energy source relates with batteries, fuel cells, capacitors, flywheels, and various hybrid sources. Thus, the process of identifying the preferred features and packaging options for electric motor drive has to be carried out at the system level. The interactions between subsystems and those likely impacts of system trade-offs must be examined.

Classification

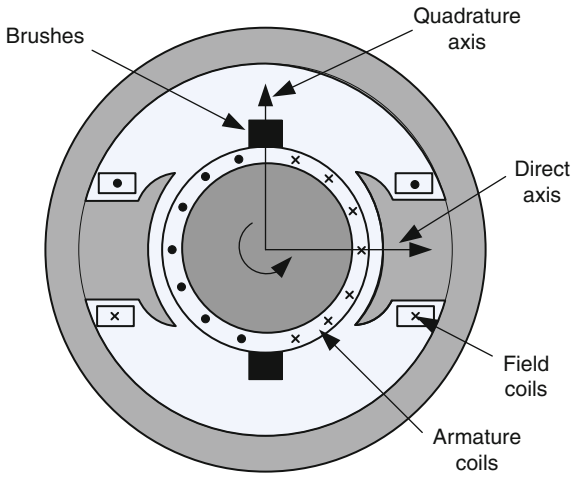
Electric motors have been available for over a century. The evolution of motors, unlike that of electronics and computer science, has been long and relatively slow. Nevertheless, the development of motors is continually fuelled by new materials, sophisticated topologies, powerful computer-aided designs (CAD), as well as modern power electronics and microelectronics. Based on the technological growth of electric motors, power electronics, microelectronics, and control strategies, more and more kinds of motor drives become available for EVs. As illustrated in Fig. 3, those traction motors applicable to EVs can be classified as two main groups, namely, the brushed motors and brushless motors. The former simply denote that they generally consist of the commutator and brushes, mainly traditional DC motors, while the latter have no brushes.

DC Motor

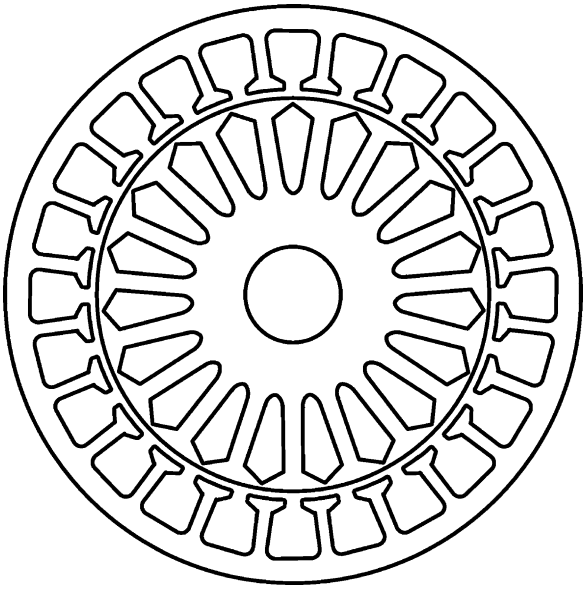
Traditionally, DC brushed motors have been loosely named as DC motors. There are typically four types of wound-field DC motors, depending on the mutual interconnection between the field and armature windings, namely, separately excited, shunt excited, series excited, and compound excited. By replacing the field winding of DC motors with PMs, PMDC motors permit a considerable reduction in stator diameter due to the efficient use of radial space. Owing to the low permeability of PMs, armature reaction is usually reduced and commutation is improved. The control principle of DC motor is simple because of the orthogonal disposition of field and armature mmfs. Figure 4 illustrates the cross section of a wound-field DC motor.



Vehicle Traction Motors. Figure 3
Classification of traction motors for EVs



Vehicle Traction Motors. Figure 4
Schematic of DC motor



Vehicle Traction Motors. Figure 5
Induction motor with squirrel cage

However, the principal problem of DC motors, due to their commutators and brushes, makes them less reliable and unsuitable for maintenance-free operation. Nevertheless, because of mature technology and simple control, DC motors have ever been prominent in electric propulsion. Actually, various types of DC motors, including series, shunt, separately excited, and PM excited, have ever been adopted by recent EVs.

Recently, technological developments have pushed brushless motors to a new era, leading to take the advantages of higher efficiency, higher power density, lower operating cost, more reliable, and maintenance free over DC brushed motors. As high reliability and maintenance-free operation are prime considerations for electric propulsion in EVs, brushless motors are becoming attractive.

Induction Motor

Induction motors (IM) are a widely accepted brushless motor type for EV traction because of their low cost, high reliability, and free from maintenance. There are two types of induction motors, namely, wound rotor and squirrel cage motors. The wound-rotor IMs are less attractive than their squirrel-cage counterparts due to their higher cost, more maintenance, and lack of sturdiness. The most common types of induction motor rotors are the squirrel cage in which aluminum bars are cast into slots in the outer periphery of the rotor, as shown in Fig. 5 [4, 5].

The main advantages of IM include: (1) Robust structure and relatively low cost; (2) Light weight, small volume, and high efficiency. The disadvantages include: (1) The limited constant-power range (only two to three times the base speed); (2) Relatively difficult control schemes due to the variable equivalent parameters.

Conventional control of induction motors such as variable-voltage variable frequency (VVVF) cannot provide the desired performance. One major reason is due to the nonlinearity of their dynamic model. With the advent of microcomputer era, the principle of field-oriented control (FOC) of induction motors has been accepted to overcome their control complexity due to the nonlinearity. Notice that FOC is also known as vector control or decoupling control. Nevertheless, these EV induction motors employing FOC still suffer from low efficiency at light loads and limited constant-power operating region. On the one hand, an online efficiency-optimizing control scheme has been developed for these EV induction motors [6], which can reduce the consumed energy by about 10% and increase the regenerative energy by about 4%, leading to extend the driving range of EVs by more than 14%. On the other hand, an electrically pole changing scheme has been

developed for EV induction motors [7, 8], which can significantly extend the constant-power operating region to over four times the base speed.

Permanent Magnet Brushless Motors

Permanent magnet brushless motors (PMBM) include sinusoidal and trapezoidal back-EMF machines. From the control schemes, they are divided into brushless DC (BLDC) and brushless AC (BLAC) motors. Generally, a trapezoidal back-EMF waveform in BLDC or a sinusoidal back-EMF waveform in BLAC is needed so as to achieve high-torque density and low-torque pulsation. The PM brushless AC motor with sinusoidal back-EMF is also called as the PM synchronous motor. The most obvious advantage of these motors is the removal of brushes, leading to eliminate many problems associated with brushes.

The PM BLAC motor can run from a sinusoidal or PWM supply without electronic commutation. When PMs are mounted on the rotor surface, they behave as non-salient synchronous motors because the permeability of PMs is similar to that of air. By burying those PMs inside the magnetic circuit of the rotor, the saliency causes an additional reluctance torque which leads to facilitate a wider speed range at constant-power operation. Similar to induction motors, those PM synchronous motors usually employ FOC or DTC for high-performance applications. Because of their inherent high power density and high efficiency, they have been accepted to have great potential to compete with induction motors for EV applications. To achieve optimal efficiency throughout the operating region, a self-tuning control has been developed for PM synchronous motors [9].

The PM BLDC motor has surface-mounted magnets on the rotor, and a concentrated fractional stator winding, which results in a low copper loss. Different from PM synchronous motors, these PM BLDC motors generally operate with shaft position sensors. Recently, a phase-decoupling PM BLDC motor has been developed for EVs, which offers the merits of outstanding power density, no cogging torque, and excellent dynamic performance [10]. Also, it can adopt advanced conduction angle control to greatly extend the constant-power operating range [11].

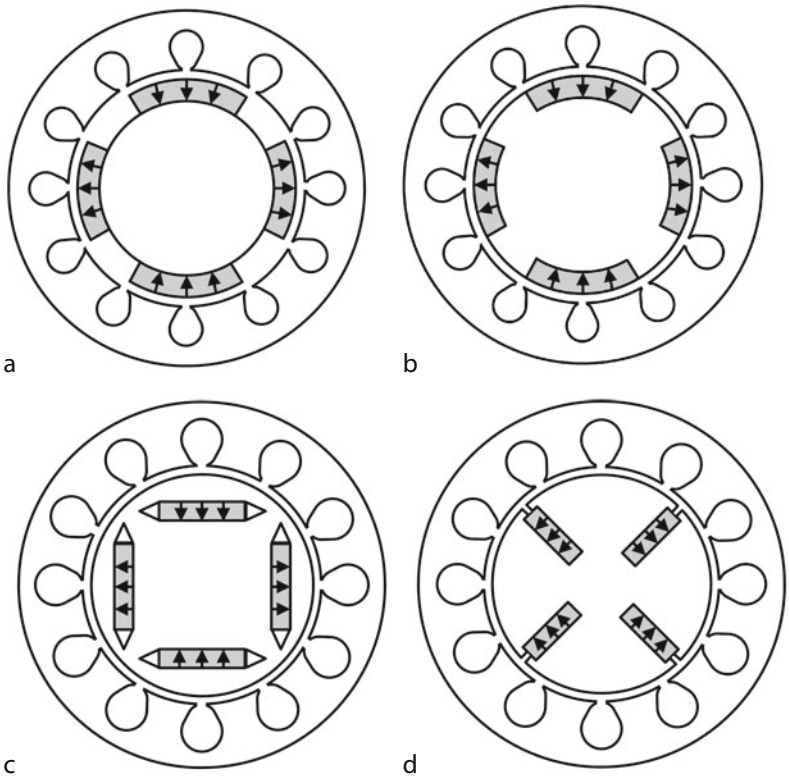
The main advantages of PM brushless motors are: (1) Light weight, small volume, and high power density as the magnetic field is excited by high-energy PMs; (2) High efficiency, high reliability, and good heat dissipation.

The main disadvantages include: (1) A comparatively narrow range of constant-power operation due to the difficultly in weakening the air-gap flux. By using some new schemes, the speed range can reach three times the base speed. However, the PM may suffer from demagnetization and possible fault. (2) Relatively high cost due to PM materials, especially in high power application [5].

Figure 6 illustrates the typical topologies of the PM brushless motors.

It should be emphasized that all the PM machines mentioned above have the magnets located in the rotor, and are referred as "rotor-PM machines," which are predominated in EV applications due to their outstanding advantages. However, the magnets usually need to be protected from the centrifugal force by employing a retaining sleeve, which is made of either stainless steel or non-metallic fiber. The rotor temperature rise may be a problem due to poor thermal dissipation, which may cause irreversible demagnetization of magnets and ultimately limit the power density of the machine. Recently, in contrast, a new type of PM machines having magnets in stator, nominated as "stator-PM machines," have reemerged and developed, which can overcome the problems suffered by rotor-PM counterparts [12]. Conceptually, the stator-PM machines employ the polarized reluctance principle, in which torque and emfs are resultant from the flux-switching action of rotor saliencies on a unipolar flux produced by PMs in the stator. Since the rotor has neither PMs nor windings, these stator-PM machines are mechanically simple and robust, hence very suitable for high-speed operation. Compared with conventional rotor-PM brushless machine topologies, generally, it is easier to limit the temperature rise of the magnets as heat is dissipated more effectively from the stator. According to the location of the PMs in stator, they can be classified as the doubly salient PM (DSPM) machine [13, 14], flux-reversal PM (FRPM) machine [15, 16], and flux-switching PM (FSPM) machine [17–19].

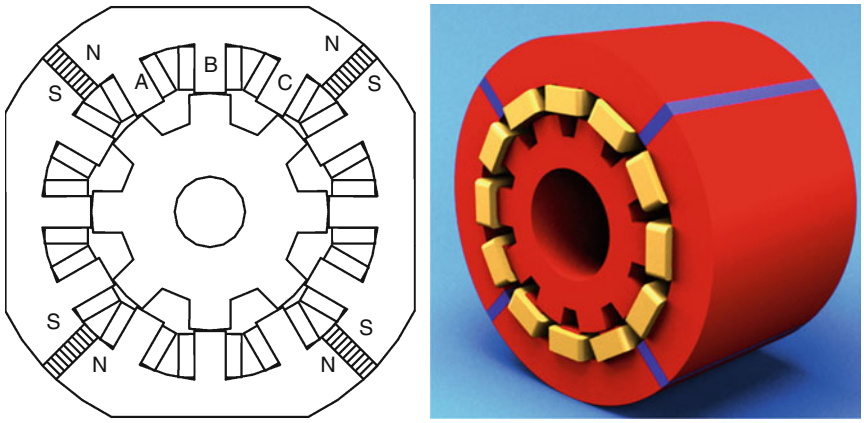
(a) Doubly Salient PM Machine: In this DSPM machine, the PMs are placed in stator back-iron.



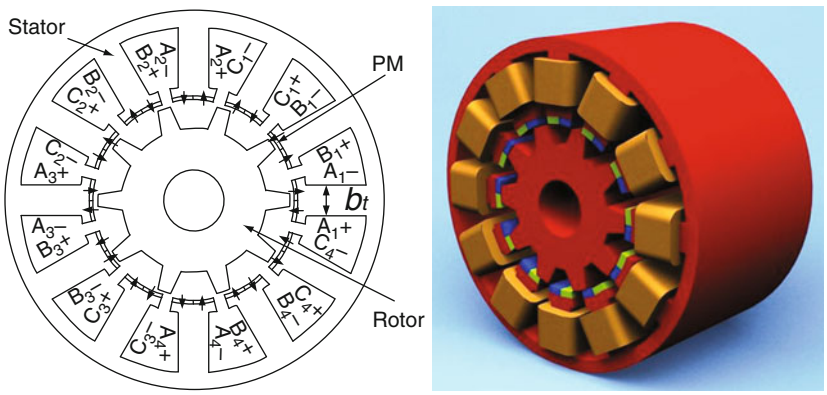
Vehicle Traction Motors. Figure 6
Typical topologies of PM brushless motors. (**a**) Surface mounted; (**b**) surface inset; (**c**) interior radial; (**d**) interior circumferential

Figure 7 shows a 12/8-pole DSPM machine topology (with 12 stator poles and 8 rotor poles). For a three-phase machine a magnet is required in the stator back-iron for every three teeth, while for a four-phase machine a magnet is required for every four teeth. The variation of the flux-linkage with each coil as the rotor rotates is unipolar, while the back-EMF waveform tends to be trapezoidal [12]. Thus, this topology is more suitable for BLDC operation. However, a major disadvantage of the DSPM motor is relatively poor torque density as compared to that of other PM brushless machines [20] due to the unipolar flux-linkage, although, as reported in [14], it can still be higher than that of an induction machine.

(b) Flux-Reversal PM Machine: The FRPM machine has the magnets located on the surface of stator teeth and concentrated windings. Figure 8 illustrates a 12/10-pole FRPM machine topology. Each stator tooth has a pair of magnets of different polarity mounted at its surface. When a coil is excited, the field under one magnet is reduced while that under the other is increased, and the salient rotor pole rotates toward the stronger magnetic field. The flux-linkage with each coil reverses polarity as the rotor rotates. Thus, the phase flux-linkage variation is bipolar, while the phase back-EMF waveform is, again, essentially trapezoidal. Thus, it is suitable for BLDC operation mode. Additionally, it is found that the FRPM machine exhibits fault-tolerance capability due to its natural isolation between the phases, and the variation of inductances versus rotor position is small. Such a machine topology exhibits a low winding inductance, while the magnets are more vulnerable to partial irreversible demagnetization. In addition, significant eddy-current loss may be



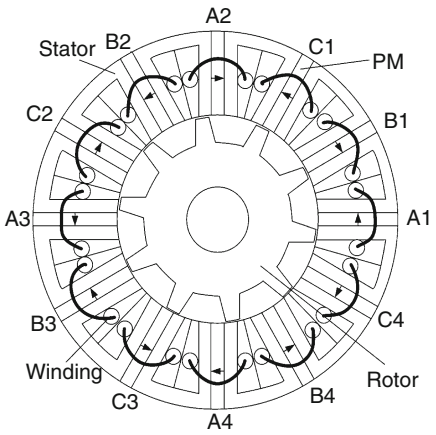
Vehicle Traction Motors. Figure 7
A 12/8-pole DSPM machine

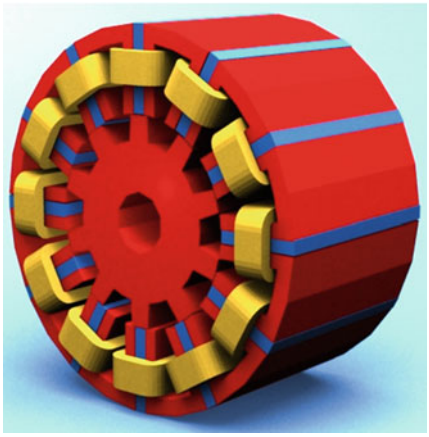


Vehicle Traction Motors. Figure 8
FRPM machine

induced in the magnets, which also experience a significant radial magnetic force. Further, since the air-gap flux density is limited by the magnet remanence, the torque density may be compromised [2].

(c) Flux-Switching PM Machine: In this FSPM machine, the stator consists of U-shaped laminated segments between which circumferentially magnetized PMs are sandwiched, while the direction of magnetization is being reversed from one magnet to the next. Figure 9 shows a 12/10-pole FSPM machine topology. Each stator tooth comprises two adjacent laminated segments and a PM. Thus, flux-concentration can be readily incorporated, so that low-cost ferrite magnets can be employed [2]. In addition, in contrast to conventional PM brushless machines, the influence of the armature reaction field on the working point of the magnets is minimal. As a consequence, the electric loading of FSPM machines can be very high. Therefore, since the phase flux-linkage waveform is bipolar, the torque capability is significantly higher than that of a DSPM machine [20]. Due to the magnetic reluctance difference between the two pairs of coils composing a phase, the resultant phase emf waveforms are essentially sinusoidal without any additional measures [18], which makes them more appropriate for BLAC





Vehicle Traction Motors. Figure 9
FSPM machine

operation. In addition, since a high per unit winding inductance can readily be achieved, such machines are eminently suitable for constant-power operation over a wide speed range.

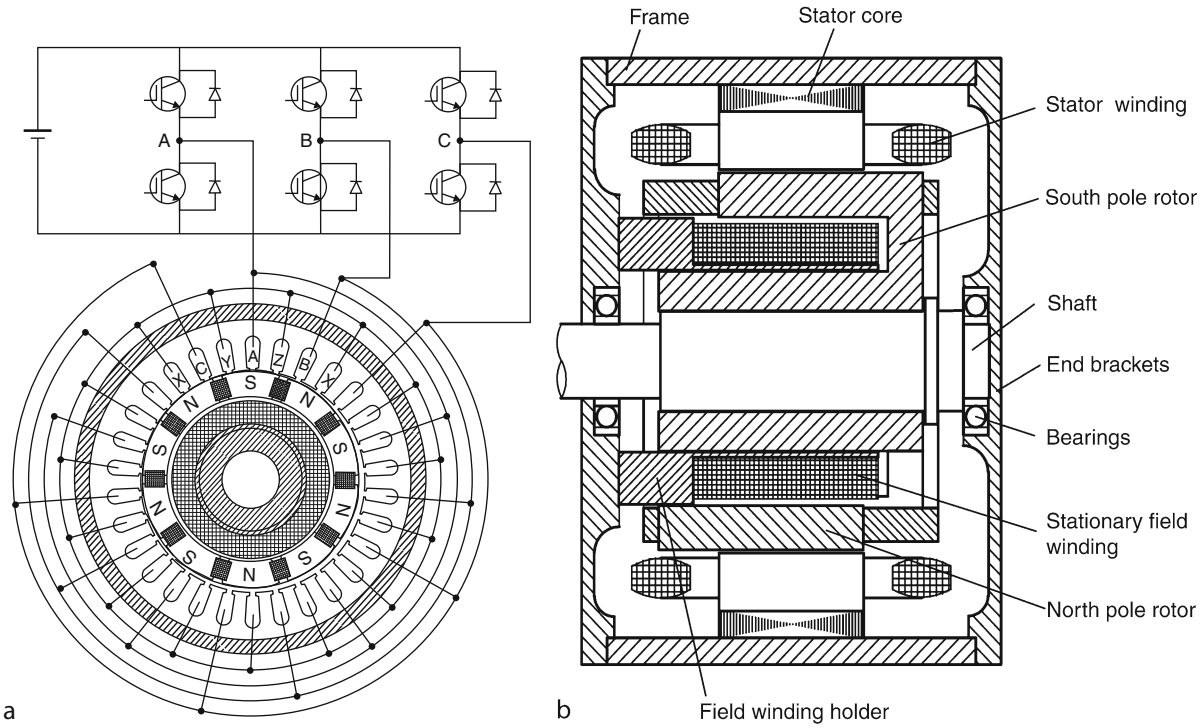
PM Hybrid Motor

Although the PM brushless motors possess the highest efficiency and power density over the others, they suffer from a difficulty in flux control. Hence, the current phase angle has to be progressively advanced as the speed is increased above the base-speed so that a demagnetizing d-axis current component is produced which reduces the flux-linkage. Ultimately, however, this may cause partial irreversible demagnetization of the magnets. At the same time, due to the inverter voltage and current limits, the torque-producing q-axis current component has to be reduced correspondingly. Consequently, the torque and power capabilities are limited [2]. Thus, a compromise has to be made between the low-speed torque capability and high-speed power capability. Hybrid PM and field current excitation have been shown to be beneficial in improving the power capability in the extended speed range, enhancing the low-speed torque capability, and improving the overall operational efficiency. Figures 10 and 11 show PM hybrid motors with rotary and stationary PMs, respectively [10, 21]. The PM hybrid motor is a special type of PM brushless motors. In this motor, an auxiliary DC field winding is so incorporated that the air-gap flux is a resultant of the PM flux and field-winding flux. These PM hybrid motors offer many attractive features due to the presence of the hybrid PM field [3]:

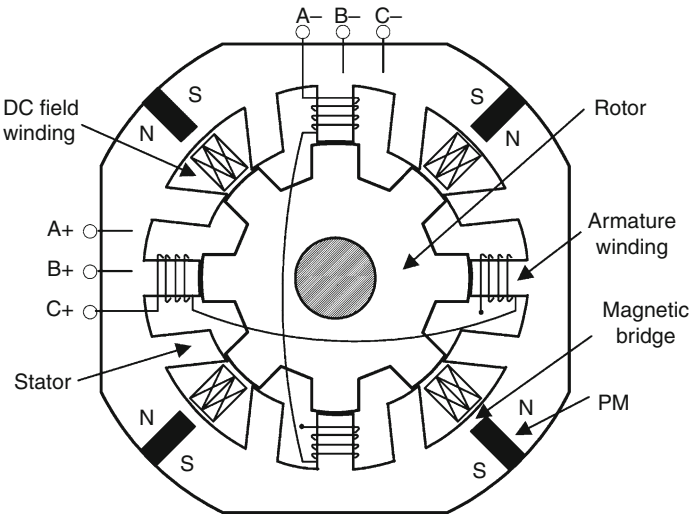
1. By varying the polarity and magnitude of the DC field current, the air-gap flux density becomes easily controllable.
2. By realizing flux strengthening, the machine can offer the exceptionally high-torque feature, which is very essential for cold cranking HEVs or providing temporary power for vehicular overtaking and hill climbing.
3. By realizing flux weakening, the machine can offer the exceptionally wide speed constant-power feature, which is very essential for EV cruising.
4. By online tuning the air-gap flux density, the machine can maintain a constant voltage output under generation or regeneration over a very wide speed range, which is very essential for battery charging of various EVs.
5. By online tuning the air-gap flux density, the machine can also offer efficiency-optimizing-control (EOC), which is highly desirable for EVs.

Switched Reluctance Motor

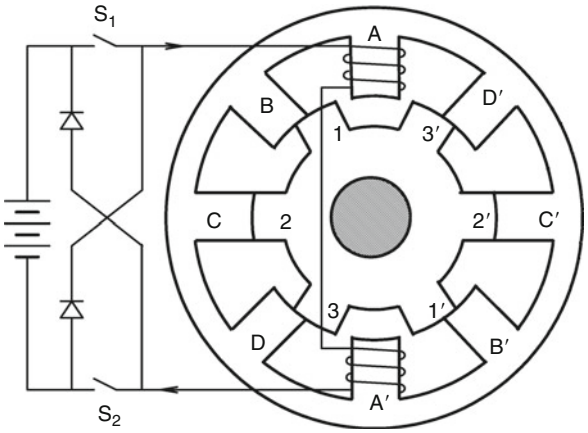
SR motors have been recognized to have considerable potential for EV applications. Figure 12 shows the schematic of an 8/6-pole SR motor. SR motors have the definite advantages of simple construction, low manufacturing cost, inherent fault tolerance, and outstanding torque-speed characteristics for EV



Vehicle Traction Motors. Figure 10
Hybrid PM machine with rotary PMs



Vehicle Traction Motors. Figure 11
Hybrid PM machine with stationary PMs



Vehicle Traction Motors. Figure 12
Basic structure of switched reluctance motor drive (only one phase winding shown)

Vehicle Traction Motors. Table 1 Applications of EV motors

EV models	EV motors
Fiat Panda Elettra	Series DC motor
Mazda Bongo	Shunt DC motor
Conceptor G-Van	Separately excited DC motor
Suzuki senior tricycle	PMDC motor
Fiat Seicento Elettra	Induction motor
Ford Th!nk City	Induction motor
GM EV1	Induction motor
Honda EV Plus	PM synchronous motor
Nissan Altra	PM synchronous motor
Toyota RAV4	PM synchronous motor
Chloride Lucas	Switched reluctance motor
Toyota Prius (2005)	PM BLDC motor
Honda Civic	PM BLDC motor

propulsion. Although they possess the simplicity in construction, it does not imply any simplicity of their design and control. Because of the heavy saturation of pole tips and the fringe effect of poles and slots, their design and control are difficult and subtle. Also, they usually exhibit relatively high acoustic noise, vibration, and torque ripple problems. Recently, an optimum design approach to SR motors has been developed [22], which employs finite-element analysis to minimize the total motor losses while taking into account the constraints of pole arc, height, and maximum flux density. Also, fuzzy sliding mode control has been developed for those EV SR motors so as to handle the motor nonlinearities and minimize the control chattering [23, 24].

The motor types that have ever been adopted by recent EVs are indicated by shaded blocks in Fig. 3. Table 1 also illustrates their recent applications to flagship EVs.

In order to evaluate the aforementioned EV motor types, a point grading system is adopted. The grading system consists of six major characteristics and each of them is graded from one to five points. As listed in Table 2, this evaluation indicates that induction motors and PM brushless motors are relatively most acceptable. When the cost of PM material has significant improvements, the PM brushless (including AC or DC) motors will be most attractive. Conventional DC motors seem to be losing their competitive edges, whereas both SR and PM hybrid motors have increasing potentials for EV propulsion.

Design Consideration

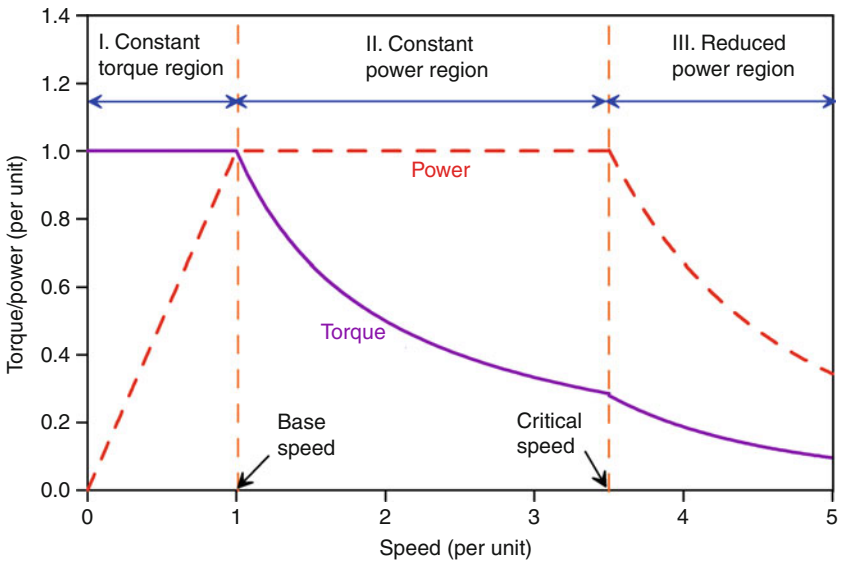
Basic Consideration

The basic consideration of motor design includes magnetic loading – the peak of fundamental component of radial flux density in the air-gap of the motor, electric loading – the total rms current per unit length of periphery of the motor or ampere-turns per unit periphery, power per unit volume and weight, torque per unit volume and weight, flux density at each part of the magnetic circuit, speed, torque and power, losses and efficiency, and thermal design and cooling.

The corresponding key issues are better utilization of steel, magnet, and copper; better electromagnetic coupling; better geometry and topology; better thermal design and cooling; understanding the limits on the motor performance; and understanding the relationship among geometry, dimensions, parameters, and performance, thus to achieve higher power per unit weight, higher torque per unit weight, and better performance.

Vehicle Traction Motors. Table 2 Evaluation of EV motors

	DC motor	Induction motor	PM brushless motor	SR motor	PM hybrid motor
Power density	2.5	3.5	5	3.5	4
Efficiency	2.5	3.5	5	3.5	5
Controllability	5	4	4	4	4.5
Reliability	3	5	4	5	4
Maturity	5	5	5	4	3
Cost	4	5	3	4	3
Total	22	26	26	24	23.5



Vehicle Traction Motors. Figure 13
Ideal torque/power-speed characteristics

Traction motor drives for EVs should be designed, as close as possible, to the ideal torque/power-speed characteristics as shown in Fig. 13. In the constant-torque region I, the maximum torque capability is determined by the current rating of the inverter, while in the constant-power region II, flux weakening or commutation phase advance has to be employed due to the inverter voltage and current limits. In region III, the torque and power are reduced due to the increasing influence of the back-emf. However, the power capability and the maximum speed can be enhanced without sacrificing the low-speed torque capability by employing a DC–DC voltage booster [2], a technique which is employed in the Toyota hybrid system, or by employing series/parallel winding connections, i.e., series connection at low speed and parallel connection at high speed, as demonstrated in [25] and [26].

Electrical machine design cannot be undertaken in isolation, but must account for the control strategy and the application requirements, both static and dynamic. Hence, a system-level design approach is essential.

System Consideration

Vehicle operation consists of three main segments. They are: (1) the initial acceleration; (2) cruising at

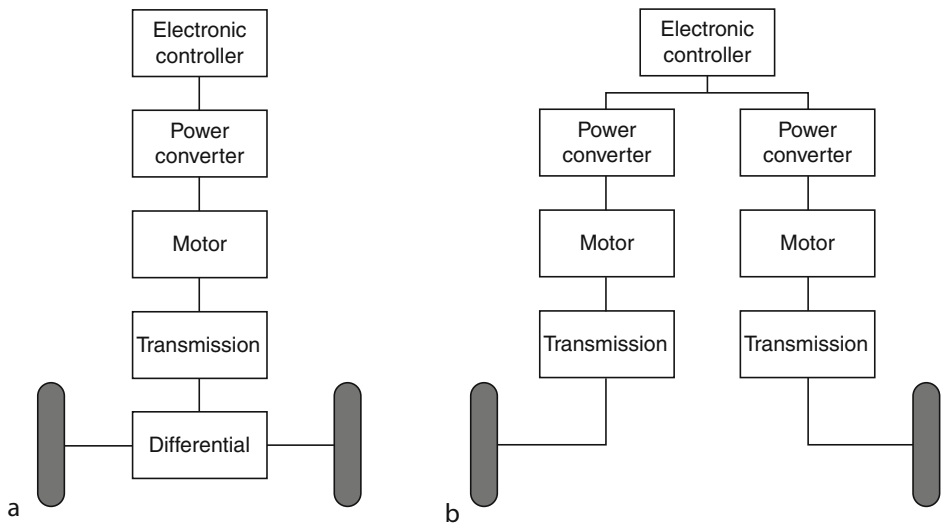
vehicle rated speed; and (3) cruising at the maximum speed. These three operations provide the basic design constraints for the EV and HEV drive train.

Apart from satisfying the aforementioned special requirements, the design of traction motors also depends on the system technology of EVs. From the technological point of view, the following key issues should be considered:

1. Single- or multiple-motor configurations. One adopts a single motor to propel the driving wheels, while another uses multiple motors permanently coupled to individual driving wheels. The single-motor configuration has the merit of using only one motor which can minimize the corresponding size, weight, and cost. On the other hand, the multiple-motor configuration takes the advantages to reduce the current/power ratings of individual motors and evenly distribute the total motor size and weight. Also, the multiple-motor one needs additional precaution to allow for fault tolerance during the electronic differential action. The functional block diagrams of single- and dual-motor configurations are shown in Fig. 14, while their comparison is listed in Table 3. Since these two configurations have their individual merits, both of them have been employed by modern EVs. For example, the single-motor configuration has been adopted in the GM EV1 while the dual-motor configuration has been adopted in the NIES Luciole. Nevertheless, as reliability is of utmost importance for EVs, the use of single-motor configuration is rekindling, especially for commercialization.
2. Fixed- or variable-gearing transmissions. It is also classified as single-speed and multiple-speed transmissions. The former adopts single-speed fixed-gearing, while the latter uses multiple-speed variable gearing together with the gearbox and clutch. Based on fixed-gearing transmission, the motor should be so designed that it can provide both high instantaneous torque (three to five times the

Vehicle Traction Motors. Table 3 Comparison of single- and dual-motor configurations

	Single motor	Dual motor
Cost	Lower	Higher
Size	Lumped	Distributed
Weight	Lumped	Distributed
Efficiency	Lower	Higher
Differential	Mechanical	Electronic
Reliability	Higher	Lower
Failure modes	Better	Worse



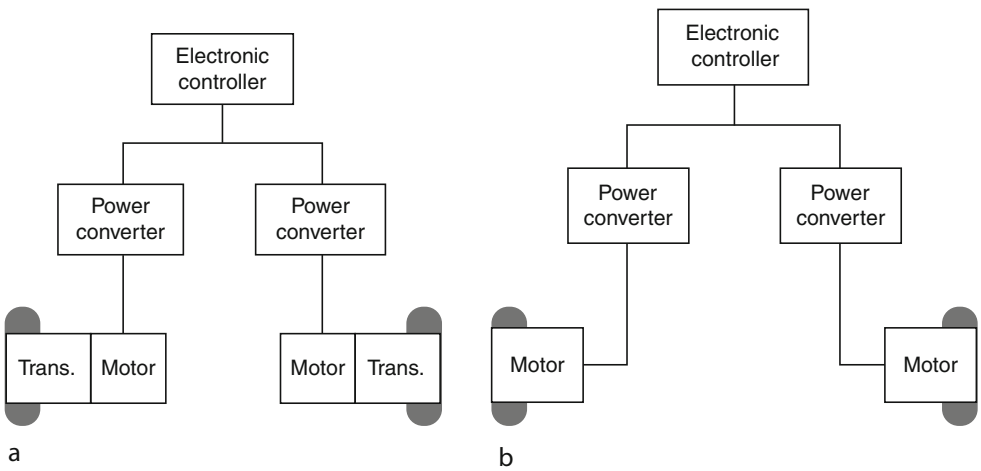
Vehicle Traction Motors. Figure 14
(**a**) Single-motor and (**b**) dual-motor configurations

rated value) in the constant-torque region and high operating speed (three to five times the base speed) in the constant-power region. On the other hand, the variable-gearing transmission provides the advantage of using conventional motors to achieve high starting torque at low gear and high cruising speed at high gear. However, there are many drawbacks on the use of variable gearing such as the heavy weight, bulky size, high cost, less reliable, and more complex. Table 4 gives a comparison of fixed-gearing and variable-gearing transmissions. Actually, almost all the modern EVs adopt fixed-gearing transmission.

3. Geared or gearless. The use of fixed-speed gearing with a high gear ratio allows EV motors to be designed for high-speed operation, resulting high power density. The maximum speed is limited by the friction and windage losses as well as transaxle tolerance. On the other hand, EV motors can directly drive the transmission axles or adopt the in-wheel drive without using any gearing (gearless operation). However, it results the use of low-speed outer-rotor motors which generally suffer from relatively low-power density. The breakeven point is whether this increase in motor size and weight can be outweighed by the reduction of gearing. Otherwise, the additional size and weight will cause suspension problems in EVs. The functional block diagrams of geared and gearless in-wheel motor configurations are shown in Fig. 15. Both of them have been employed by modern EVs. For examples, the high-speed geared inner-rotor in-wheel motor has been adopted in the NIES Luciole while the low-speed gearless outer-rotor in-wheel motor was adopted in the TEPCO IZA. Nevertheless, with the advent of compact planetary gearing, the use of high-speed planetary-geared in-wheel motors is becoming more attractive than the use of low-speed gearless in-wheel motors.
4. System voltage. The design of traction motors is greatly influenced by the voltage level of the EV system. Reasonable high-voltage motor design can be adopted to reduce the cost and size of inverters. If the desired voltage is too high, a large number of

Vehicle Traction Motors. Table 4 Comparison of fixed- and variable-gearing transmissions

	Fixed gearing	Variable gearing
Motor rating	Higher	Lower
Inverter rating	Higher	Lower
Cost	Lower	Higher
Size	Smaller	Larger
Weight	Lower	Higher
Efficiency	Higher	Lower
Reliability	Higher	Lower



Vehicle Traction Motors. Figure 15
In-wheel motor configurations. (**a**) Geared motor; (**b**) Gearless motor

batteries will be connected in series, leading to the reduction of interior and luggage spaces, the increase in vehicle weight and cost, as well as the degradation of vehicle performances. Since different EV types adopt different system voltage levels, the design of EV motors needs to cater for different EVs. Roughly, the system voltage is governed by the battery weight which is about 30% of the total vehicle weight. In practice, higher power motors adopt higher voltage levels. For examples, the GM EV1 adopts the 312-V voltage level for its 102-kW motor, whereas the Reva EV adopts the 48-V voltage level for its 13-kW motor.

5. Integration. The integration of the motor with the converter, controller, transmission and energy source is prime important consideration. The EV motor designer should fully understand the characters of these components, thus to design the motor under these given environments. It is quite different with the normal standard motors under standard power source for normal industrial drives.

Efficiency

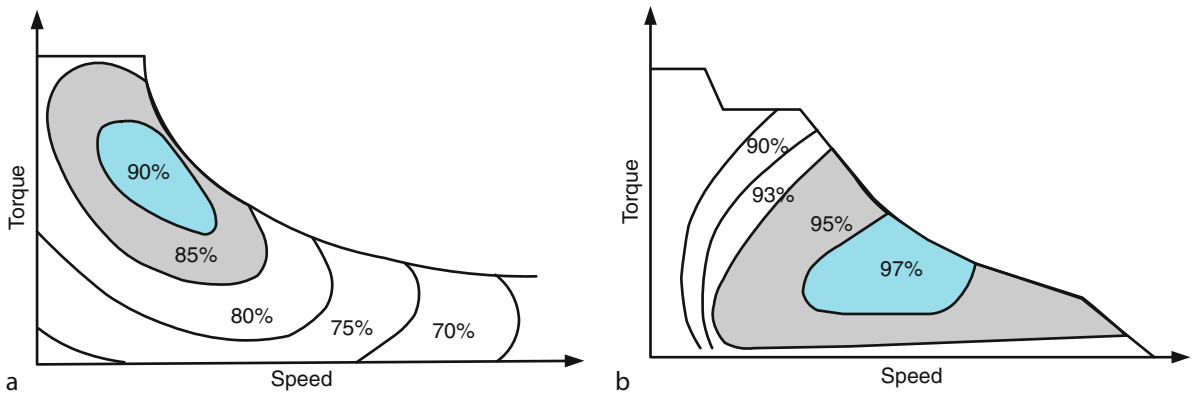
The efficiency may be classified into two types, namely, energy efficiency and power efficiency. The energy efficiency η_e is the ratio of energy output to energy input, while the power efficiency η_p is the ratio of power output to power input. So, they can simply be expressed as:

$$\eta_e = \frac{E_{out}}{E_{in}}$$

$$\eta_p = \frac{P_{out}}{P_{in}}$$

For industrial operation, these two efficiencies may not be necessarily distinguishable. On the contrary, for vehicular operation, there is a significant difference because the power efficiency varies continually during the operation of most vehicles. Thus, it is necessary to delineate the power efficiency associated with the speed and torque conditions. Instead of using a particular operating point (such as rated power at rated torque and rated speed) to describe the power efficiency of a vehicle subsystem or component, an efficiency map is generally adopted. Figure 16 shows typical efficiency maps of a three-phase induction motor and a PM BLDC motor for propelling an EV. Hence, the energy efficiency can be derived by summing powers over a given time period.

Regenerative braking is a definite advantage of EVs over internal combustion engine vehicles (ICEVs). During braking, the motor operates in the regenerative mode which converts the reduction of kinetic energy during braking into electrical energy, hence recharging the batteries. On average, the amount of convertible energy is only about 30–50% as there is significant dissipation in road load. Assuming that the in/out efficiency of the drivetrain and energy source is about 70%, the amount of energy actually stored in the batteries is about 21–35%. This is known as the regenerative braking efficiency. Similar to the previous case, in order to depict the value at different loads, a regenerative braking efficiency map should be adopted.

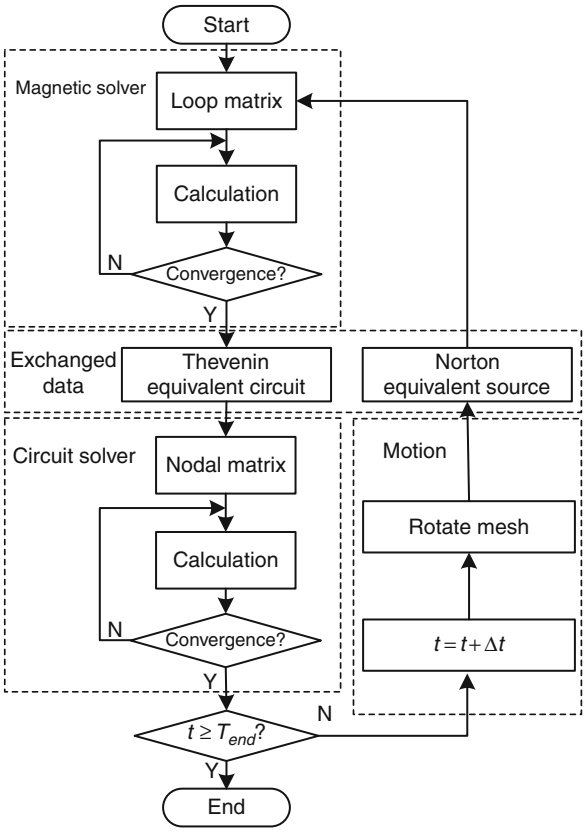


Vehicle Traction Motors. Figure 16
Typical power efficiency maps of EV traction motors. (**a**) Induction motor; (**b**) PM BLDC motor

Design Methodology

To keep up with the stringent requirement and fast changing motor topologies, the design of motors has turned to CAD. Basically, there are two major design approaches – circuit and field. In essence, the circuit approach is based on equivalent circuit analysis while the field approach depends on electromagnetic field analysis. The field approach takes the advantages of more accurate results, greater knowledge of the critical areas, as well as capabilities of handling complicated machine geometry and nonlinear materials. Recently, the finite-element method (FEM) has been regarded as one of the most powerful tools for electromagnetic field analysis of EV motors. The FEM outranks other numerical methods because of its flexibility and applicability in stress and thermal field analyzes.

The performance of a motor drive depends on not only the motor, but also the control circuits. And there are strong coupling between the magnetic circuit and electric circuit of a motor drive. Traditionally, magnetic circuit and electric circuit of electric machines, however, are separately dealt with in dynamic simulations. To evaluate accurately the performance of a motor drive in design procedure, the co-simulation method for the motor drive may be required. In co-simulation, the magnetic circuit and the electric circuit are coupled in time-domain, providing the possibility of system-level simulation. Finite-element time-domain modeling coupled with the equations of circuit and motion is an accurate and detailed approach of simulating drive system performance. Figure 17 illustrates the field-circuit-motion coupled method. The modeling tools for the co-simulation consist of two separate packages, namely, the magnetic solver, Maxwell 2D® and the circuit solver, Simplorer®. The magnetic solver calculates two-dimensional transient magnetic problem of motor drive, while electric circuit and controller are supplied by the circuit solver. Incorporating electric circuit equations into equations of the finite-element system, the magnetic solver uses a loop form of the magnetic equations. At the same time, the circuit solver uses a nodal form of the circuit equations. In each time step, the circuit solver forms a Norton equivalent source of the drive circuit at the coupling pins between the motor and the rest of the drive system.



Vehicle Traction Motors. Figure 17
Flowchart of a co-simulation method

The magnetic solver converts it to a loop matrix and solves the finite-element equations. Finally, the magnetic solver outputs a Thevenin equivalent circuit for the next time step of the circuit solver. This parameter-based coupling enhances the solution accuracy and stability. When the simulation runs in the circuit solver, the magnetic solver will start automatically in a co-simulation mode. At each co-simulation time step, both the simulators exchange the calculated data, and results achieved by one solver will be exported to the other solver in the next step. The co-simulation model allows an easy access to all available components such as linear or nonlinear resistances, capacitances, inductances, various diodes, controlled switches, independent sources, voltages, and current probes [27, 28].

Control Consideration

Power Electronics

EV Power Devices In the past decades, power semiconductor device technology has made tremendous progress. These power devices have grown in power rating and performance by an evolutionary process. Among existing power devices, the power diode behaves as an uncontrolled switch, whereas the others, including the thyristor, gate turnoff thyristor (GTO), power bipolar-junction transistor (BJT), power metal-oxide field-effect transistor (MOSFET), insulated-gate bipolar transistor (IGBT), static-induction transistor (SIT), static-induction thyristor (SITH), and MOS-controlled thyristor (MCT), are externally controllable. Active research is still being pursued on the development of high performance power devices.

Before selecting the preferred power devices for electric propulsion, the following requirements have to be considered:

- Ratings. The voltage rating is based on the battery nominal voltage, maximum voltage during charging, and maximum voltage during regenerative braking. On the other hand, the current rating depends on the motor peak power rating and number of power devices connected in parallel. When paralleling these devices, on-state and switching characteristics have to be matched.
- Switching frequency. Switching at higher frequencies can bring down the filter size and help to meet the electromagnetic interference (EMI) limitation requirements. Over the switching frequency of 20 kHz, there is no acoustic noise problem.
- Power losses. The on-state conduction drop or loss should be the minimum while the switching loss should be as low as possible. Since higher switching frequencies increase the switching loss, switching the device at about 10 kHz seems to be an optimum for efficiency, power density, acoustic noise, and EMI considerations. The leakage current should also be less than 1 mA to minimize the off-state loss.
- Base/gate driveability. The device should allow for simple and secure base/gate driving. The corresponding driving signal may be either triggering voltage/current or linear voltage/current. The voltage-mode driving involves very little energy and is generally preferable.
- Dynamic characteristics. The dynamic characteristics of the device should be good enough to allow for high dv/dt capability, high di/dt capability, and easy paralleling. The internal antiparallel diode should have similar dynamic characteristics as the main device.
- Ruggedness. The device should be rugged to withstand a specific amount of avalanche energy during overvoltage and be protected by fast semiconductor fuses during over-current. It should operate with no or minimal use of snubber circuits. Since EVs are frequently accelerated and decelerated, the device is subjected to thermal cycling at frequent intervals. It should reliably work under these conditions of thermal stress.
- Maturity and cost. Since the cost of power devices is one of the major parts in the total cost of electric propulsion systems, these devices should be economical. Some recent power devices such as the high power MCT are not yet mature for EV applications.

Taking into account the above requirements, the GTO, power BJT, power MOSFET, IGBT, and MCT are considered for electric propulsion. The thyristor is not considered because it requires additional commutating components to turn off and its switching frequency is limited to 400 Hz. The SIT and SITH are also excluded because of their normally turn-on property and limited availability. In order to evaluate their suitability, a point grading system is adopted, which consists of eight major characteristics and each of them is graded from one to five points. From Table 5, the power MOSFET, IGBT, and MCT score high points which indicate that they are particularly suitable for EV propulsion. Due to its highest score, the IGBT is almost exclusively used for modern EVs. Nevertheless, the power MOSFET has also been accepted for those relatively low-power electric tricycles and bikes.

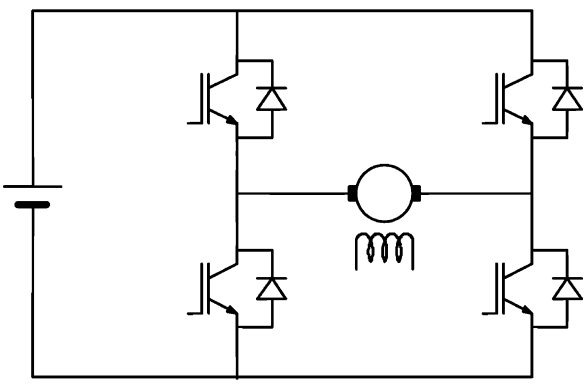
EV Power Converters The evolution of power converter topologies normally follows that of power devices, aiming to achieve high power density, high efficiency, high controllability, and high reliability [29].

Vehicle Traction Motors. Table 5 Evaluation of EV power devices

	GTO	BJT	MOSFET	IGBT	MCT
Ratings	5	4	2	5	3
Switching frequency	1	2	4	4	4
Power losses	2	3	4	4	4
Base/gate driveability	2	3	5	5	5
Dynamic characteristics	2	3	5	5	5
Ruggedness	3	3	5	5	5
Maturity	5	5	4	4	2
Cost	4	4	4	4	2
Total	24	27	33	36	30

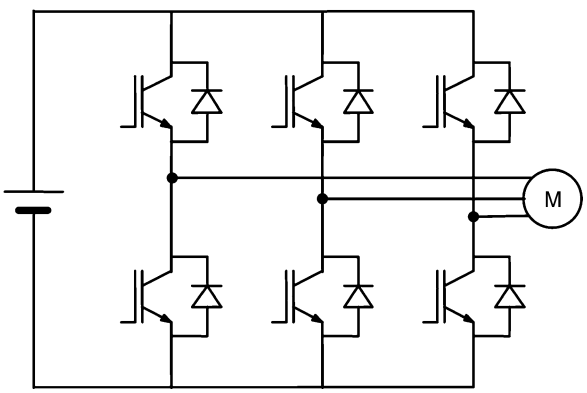
Power converters may be AC–DC, AC–AC at the same frequency, AC–AC at different frequencies, DC–DC or DC–AC. Loosely, DC–DC converters are known as DC choppers while DC–AC converters are known as inverters, which are respectively used for DC and AC motors for electric propulsion.

Initially, DC choppers were introduced in the early 1960s using force-commutated thyristors that were constrained to operate at low switching frequency. Due to the advent of fast-switching power devices, they can now be operated at tens or hundreds of kilohertz. In electric propulsion applications, two-quadrant DC choppers are desirable because they convert battery DC voltage to variable DC voltage during the motoring mode and revert the power flow during regenerative braking. Furthermore, four-quadrant DC choppers are employed for reversible and regenerative speed control of DC motors. A four-quadrant DC chopper is shown in Fig. 18.



Vehicle Traction Motors. Figure 18
Four-quadrant DC chopper

Inverters are generally classified into voltage-fed and current-fed types. Because of the need of a large series inductance to emulate a current source, current-fed inverters are seldom used for electric propulsion. In fact, voltage-fed inverters are almost exclusively used because they are very simple and can have power flow in either direction. A typical three-phase full-bridge voltage-fed inverter is shown in Fig. 19. Its output waveform may be rectangular, six-step or PWM, depending on the switching strategy for different applications. For example, a rectangular output



Vehicle Traction Motors. Figure 19
Three-phase full-bridge voltage-fed inverter

waveform is produced for a PM BLDC motor, while a six-step or PWM output waveform is for an induction motor. It should be noted that the six-step output is becoming obsolete because its amplitude cannot be directly controlled, and its harmonics are rich. On the other hand, the PWM waveform is harmonically optimal and its fundamental magnitude and frequency can be smoothly varied for speed control.

Starting from the last decade, numerous PWM switching schemes have been developed for voltage-fed inverters, focusing on the harmonic suppression, better utilization of DC voltage, tolerance of DC voltage fluctuation, as well as suitability for real-time and microcontroller-based implementation [29]. These schemes can be classified as voltage-controlled and current-controlled PWM. The state-of-the-art voltage-controlled PWM schemes are natural or sinusoidal PWM, regular or uniform PWM, harmonic elimination or optimal PWM, delta PWM, carrierless or random PWM, and equal-area PWM. On the other hand, the use of current control for voltage-fed inverters is particularly attractive for high-performance motor drives because the motor torque and flux are directly related to the controlled current. The state-of-the-art current-controlled PWM schemes are hysteresis-band or band–band PWM, instantaneous current control with voltage PWM, and space vector PWM.

Soft-Switching EV Converters Instead of using hard or stressed switching, power converters can adopt soft or relaxed switching. The key of soft switching is to employ a resonant circuit to shape the current or voltage waveform such that the power device switches at zero-current or zero-voltage condition. In general, the use of soft-switching converters possesses the following advantages:

- Due to zero-current or zero-voltage switching condition, the device switching loss is practically zero, thus giving high efficiency.
- Because of low heat sinking requirement and snubberless operation, the converter size and weight are reduced, thus giving high power density.
- The device reliability is improved because of minimum switching stress during soft switching.
- The EMI problem is less severe and the machine insulation is less stressed because of lower *dv/dt* resonant voltage pulses.
- The acoustic noise is very small because of high frequency operation.

On the other hand, their key drawbacks are the additional cost of the resonant circuit and the increased complexity. Although soft-switching DC–DC converters have been widely accepted by switched-mode power supplies, the corresponding development for EV propulsion is much slower. As the pursuit of power converters having high efficiency and high power density for EV propulsion is highly desirable, the development of EV soft-switching power converters is in progress [30–32]. Table 6 gives a comparison between hard-switching and soft-switching converters for EV propulsion.

Although there have been many soft-switching DC–DC converters developed for switched-mode power supplies, these converters cannot be directly applied to DC motors for EV propulsion. Apart from suffering excessive voltage and current stresses, they cannot handle backward power flow during regenerative braking. It should be noted that the capability of regenerative braking is very essential for EVs as it can

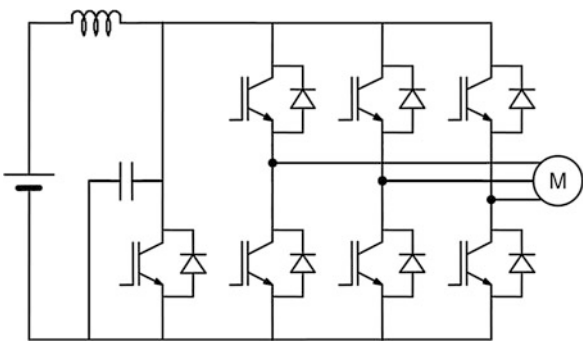
Vehicle Traction Motors. Table 6 Comparison of hard switching and soft switching for EV converters

	Hard switching	Soft switching
Switching loss	Severe	Almost zero
Overall efficiency	Norm	Possibly higher
Heat-sinking requirement	Norm	Possibly lower
Hardware count	Norm	More
Overall power density	Norm	Possibly higher
EMI problem	Severe	Low
Dv/dt problem	Severe	Low
Modulation scheme	Versatile	Limited
Maturity	Mature	Developing
Cost	Norm	Higher

extend the vehicle driving range by up to 25%. Recently, a new soft-switching DC–DC converter, having the capability of bidirectional power flow for motoring and regenerative braking as well as the minimum hardware count, has been developed for EV DC motors [33].

The development of soft-switching inverters for AC motors (including induction motors, PM brushless motors, and PM hybrid motors) has become a research direction in power electronics. Figure 20 shows a milestone of soft-switching inverters, namely, the three-phase voltage-fed resonant DC link inverter developed in 1986 [34]. Consequently, many improved soft-switching topologies have been proposed, such as the quasi-resonant DC link, series resonant DC link, parallel resonant DC link, synchronized resonant DC link, resonant transition, auxiliary resonant commutated pole, and auxiliary resonant snubber inverters. A number of development goals of soft-switching inverters for EV propulsion have been identified, namely, efficiency over 95%, power density over 3.5 W/cm^3, switching frequency over 10–20 kHz, *dv/dt* below 1,000 V/μs, zero EMI, zero failure before the end of the vehicle life, and redundant with "limp-home" mode. Recently, the delta-configured auxiliary resonant snubber version has satisfied most of these goals, and has been demonstrated to achieve an output power of 100 kW.

Compared with the development of soft-switching inverters for AC motors, the development for SR motors has been very little [35]. Recently, a new soft-switching converter, so-called the zero-voltage-transition version, has been particularly developed for SR motors [36]. This new converter possesses the advantages that all main switches and diodes can operate at zero-voltage condition, unity device voltage and current stresses, as well as wide operating range. Moreover, it offers simple circuit topology, minimum hardware count, and low cost, leading to achieve high switching frequency, high power density, and high efficiency.



Vehicle Traction Motors. Figure 20
Three-phase voltage-fed resonant DC link inverter

Microelectronics

Since the introduction of microcomputers in 1970, the microelectronics technology has gone through an intense evolution in last 5 decades. Modern microelectronic devices can generally be classified as microprocessors, microcontrollers, and digital signal processors (DSPs).

Microprocessor technology has been used to recognize the milestone of the development of microelectronics, such as the 8086, 80186, 80286, 80386, 80486, Pentium, Pentium II, Pentium III, Pentium IV, Pentium M, and so on. Microprocessors are the CPU of microcomputer systems, which decode instructions, control activities, as well as perform all arithmetic and logical computations. Unlike microprocessors, microcontrollers, such as the 8096, 80196, and 80960, include all resources (CPU, ROM or EPROM, RAM, DMA, timers, interrupt sources, A/D and D/A converters, and I/O ports) to serve as stand-alone single-chip controllers. Thus, microcontroller-based electric propulsion systems possess definite advantages of minimum hardware and compact software. Digital signal processors (DSPs), such as the TMS320C24x/LC24x, TMS320C28x, etc., include several microcontroller peripherals such as Memory, Pulse Width Modulation (PWM) generator, Analog to Digital Converters (ADC), and Event Manager module, and have the capability of high- speed computation to implement sophisticated control algorithm for high performance motor drives for electric propulsion.

By integrating microelectronic devices and power devices on the same chip (like the integration of brain and muscle), power ICs (PICs), loosely named as "smart power," aim to further reduce the cost, minimize the size, and improve the reliability. The PIC may include the power module, control, protection, communication, and cooling. The main problems in PIC synthesis are the isolation between high-voltage and

low-voltage devices as well as cooling. Nevertheless, this technology has promising applications to electric propulsion in near future. The key is the integrating and packaging.

Control Strategies

Conventional linear control such as PID can no longer satisfy the stringent requirement placed on high-performance motor drives. In recent years, many modern control strategies have been proposed. The state-of-the-art control strategies that have been proposed for motor drives are direct torque control (DTC), efficiency optimizing control (EOC), artificial intelligent control, position-sensorless control (PSC), and so on [3].

Direct Torque Control DTC is becoming attractive for EVs, particularly for those equipped with dual-motor propulsion which desires fast torque response. It does not rely on current control and depends less on parameters. For the PM BLAC drives, the DTC controls both the torque and the flux-linkage independently [37, 38]. The controller outputs provide proper voltage vectors via the inverter in such a way that these two variables are forced to predefined trajectories.

Efficiency-Optimizing Control (EOC) EOC of motor drives is highly desirable for EVs since their on-board energy storage is very limited. Different types of motor drives may employ different ways for efficiency optimization. For the rotor-PM BLAC drives, the EOC can be achieved by online tuning the input voltage or the d-axis armature current I_{2d} to minimize the total losses P_{loss} [9], [39]

$$P_{loss}(I_{2d}, T, \omega) = P_{cu}(I_{2d}, T, \omega) + P_{Fe}(I_{2d}, T, \omega)$$

where P_{Cu} is the copper loss, and P_{Fe} is the iron loss for the given torque T and speed ω. It can be found that there is a unique optimal operating point. In particular, the minimum total losses occur at a lower d-axis armature current than that of the minimum copper loss, hence illustrating that the maximum torque per ampere control cannot maximize the efficiency of the PM BLAC drives. For the hybrid PM BLAC drive incorporated with an additional DC field winding [40], the EOC can be easily achieved by tuning the polarity and magnitude of the DC field current.

Artificial Intelligent Control All artificial intelligence-based control strategies, such as fuzzy logic control, neural network control, neuro-fuzzy control, and genetic control, are classified as artificial intelligent control (AIC). Among them, the fuzzy logic control [41] and the neural network control [42] are most mature and attractive since they can effectively handle the system's nonlinearities and sensitivities to parameter variations.

Position-Sensorless Control In order to achieve high performance for EV drives, position feedback is almost mandatory. In order to get rid of the costly and bulky position encoder, position-sensorless control (PSC) is becoming attractive [43–45]. There are various PSC techniques which can be classified as motional EMF, inductance variation, and flux-linkage variation. Basically, the position information is derived by online analysis of the voltages and currents in the machine windings.

It should be noted that the PSC can be readily incorporated into other control strategies such as the EOC, the DTC, and the AIC.

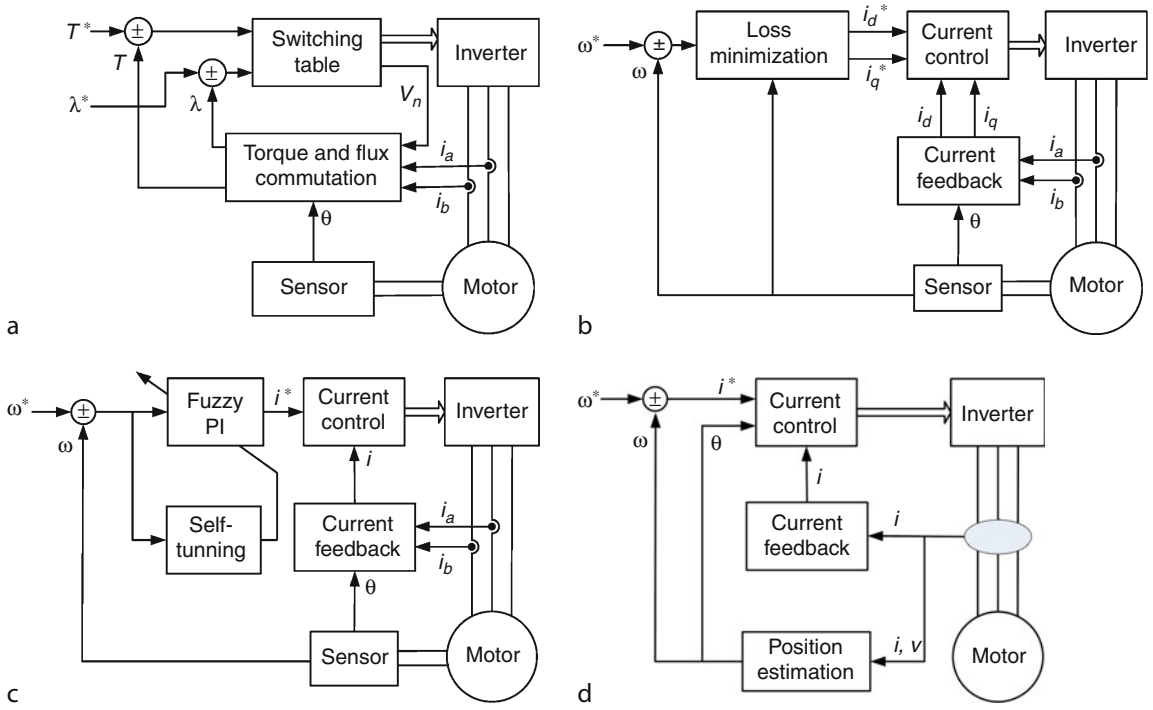
Comparison of Control Strategies As shown in Table 7, the aforementioned control strategies are compared in terms of their major advantages, major disadvantages, and typical techniques [3]. Since there are many possible strategies for the AIC, the self-tuning fuzzy PI control [41] is used for exemplification. The corresponding control block diagrams are shown in Fig. 21. Finally, some sample results of these control strategies have illustrated that the EOC can achieve the minimum total losses [39], the DTC can provide direct bang–bang control of torque [38], the AIC can achieve fast and accurate response [41], and the PSC can offer accurate estimation of rotor position [44].

Future Directions

Thanks to persistent hard work of both academic and industrial communities in the past years, the performance of traction motors for EVs has been improved greatly. With quick development of industry technology, motor drives in EVs would meet with new renovations.

Vehicle Traction Motors. Table 7 Comparison of control strategies

	Advantage	Disadvantage	Techniques
DTC	Fast torque response; no need for current control; less parameter dependence	Cause errors due to drift flux-linkage estimation and variation of stator resistance	Generate the voltage vectors using independent torque and flux computations
EOC	Minimize the overall losses; no need for accurate loss model; work for wide speed and torque range	Originate system oscillation or convergence problem	Control the input voltage or d-axis armature current; control DC field current
AIC	Flexible control algorithms; adapt nonlinearities and parameter variations	Require expert knowledge or intensive computation and sophisticated hardware	Incorporate fuzzy logic, neural network, and other AI into traditional controls
PSC	Eliminate position sensor, hence reduce system size and cost; readily merge into other controls	Require intensive computation and sophisticated hardware	Estimate the position based on motional EMF, inductance variation, or flux- linkage variation



Vehicle Traction Motors. Figure 21
Control block diagrams. (**a**) DTC; (**b**) EOC; (**c**) AIC; (**d**) PSC

The development of traction motor drives is no longer limited to the design and operation of a single motor or drive. The research trends of the traction motor drive in EVs may be concluded as follows.

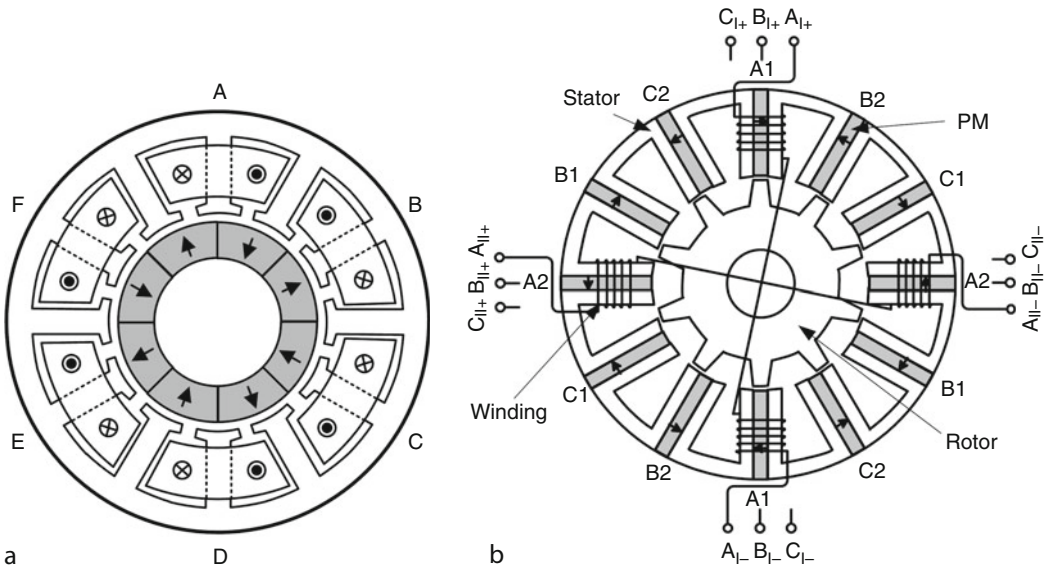
1. High-speed motors. By increasing speed, the size of electric motors may be reduced greatly, namely, higher power from smaller machines and redesigning for increased material utilization [3, 46]. Some

companies have started to focus on high speed of 16,000 r/min PM motors that can achieve field weakening within the structure of the motor and eliminate the need for a DC–DC boost converter [47].

2. System integration. It is necessary for the designers to take electrical machine, power electronics, such as converter, and energy source into consideration altogether. Furthermore, control methods will be analyzed during the machine design so as to extend the constant-power speed range, increase the starting torque, and etc.
3. Redundant and fault-tolerant motor structure. Continued operation of motor drive is an essential requirement in EV application. Therefore, the need for high degree of reliability in motor drive system has inspired much research in the area. To achieve high reliability, redundant or conservative design techniques have been employed in many motor drives. Figure 22 illustrates a 6-phase 8-pole rotor-PM motor [48] and a flux-switching motor.
4. Novel manufacture techniques. To achieve high power density, high efficiency, and low-cost motors for EVs, the manufacture technique of motors is being improved. The segmented stator and concentrated winding are examples [49]. In addition, using flat wire to replace round wire in motor windings can increase slot filling factor, enabling both a higher torque constant and a lower copper loss.
5. Novel machine topologies with composite structures and new materials. For traditional machines, each has its own merits and demerits. The composition of different machines may significantly improve the performance. Hence, the traction machines consist of different structures may be noticed in the next step.

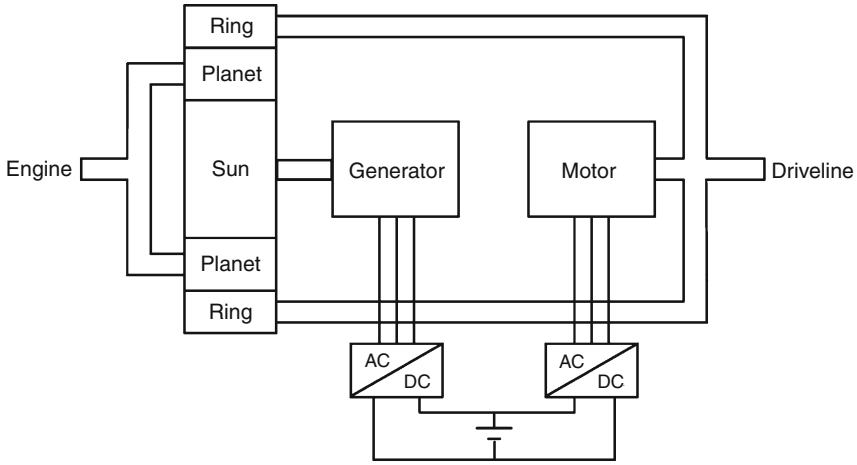
EVT Systems

In 1997, Toyota developed the first EVT system for its flagship HEV, Prius, which is a full hybrid. The schematic configuration of this EVT is shown in Fig. 23, which is mainly composed of a planetary gear, a motor, and a generator. The internal combustion engine (ICE) is attached to the planet carrier, the motor is coupled with the driveline shaft so that both are attached to the ring gear, and the generator is mounted to the sun gear [50]. By controlling the power taken by the generator and then feeding back into the motor, the ICE speed can be maintained constant when the driveline-shaft speed is varying. Thus, a continuously variable ratio between the ICE speed and the wheel speed can be achieved. Hence, this EVT system takes the following advantages.



Vehicle Traction Motors. Figure 22

Fault-tolerant motor topologies. (**a**) 6-phase 8-pole rotor-PM motor; (**b**) Flux-switching PM motor

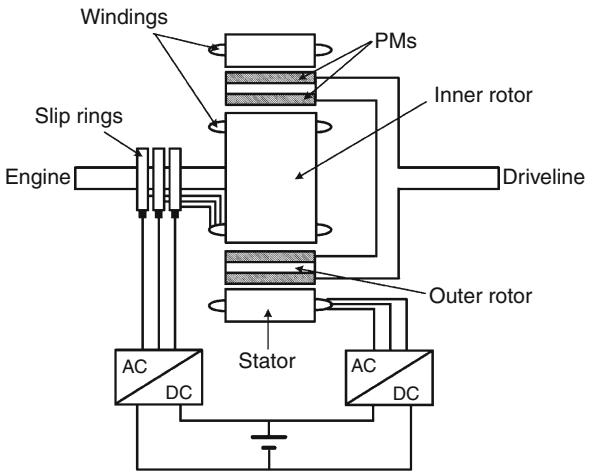


Vehicle Traction Motors. Figure 23
Planetary-geared EVT system

1. Because of the absence of clutches or shifting gears, it can significantly improve the transmission efficiency and reduce the overall size, hence increasing both the energy efficiency and the power density.
2. In the presence of continuously variable ratio between the ICE speed and the wheel speed, the ICE can always operate at its most energy-efficient operating point, hence resulting in a considerable reduction of fuel consumption.
3. The system can fully enable the idle stop, electric launch, regenerative braking, and full-throttle acceleration features, which are particularly essential for the full hybrids.

However, this planetary-geared EVT system inherits the fundamental drawbacks of planetary gearing, namely, transmission loss, gear noise, and need of regular lubrication.

In recent years, active research works have been conducted to eliminate this mechanical planetary gear while retaining the EVT propulsion. One viable approach is the use of the dual mechanical port (DMP) machine to realize power splitting for the full hybrids [51]. Figure 24 shows the EVT system integrated with the DMP machine. When installing this EVT system in a full hybrid, it offers four modes of operation, namely, cranking, charging, launching, and continuous variable transmission (CVT) [52].



Vehicle Traction Motors. Figure 24
EVT system using integrated DMP machine for HEVs

1. In the cranking mode, the battery delivers the power to crank the ICE via the primary machine until the ICE reaches the speed for ignition.
2. In the charging mode, the battery is either charged by the ICE via the inner-rotor winding when the vehicle stops motion or by the stator winding during regenerative braking.
3. In the launching mode, the battery delivers the power to launch the vehicle via the stator winding without using the ICE.

4. In the CVT mode, the input and output shafts are controlled to change the speed and the torque, respectively, so that the optimal operating line of the ICE can be achieved.

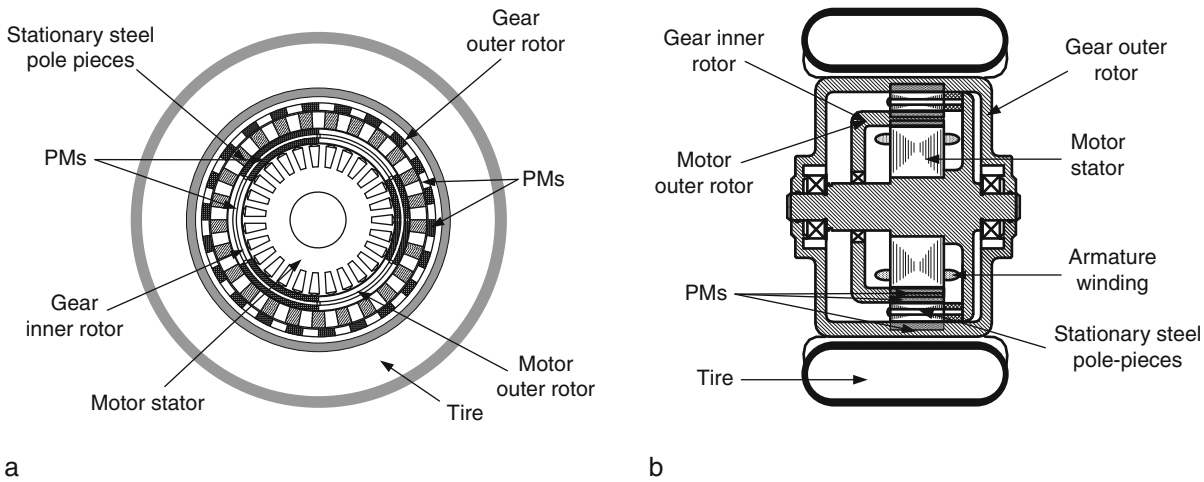
Magnetic-Geared PM Brushless Drives

For EVs, in-wheel motor drives are very attractive since they play the role of electronic differential [1]. As the wheel speed is only about 600 r/min, the in-wheel motor drive is either a low-speed gearless outer-rotor one or a high-speed planetary-geared inner-rotor one.

Although the former one takes the advantage of gearless operation, its low-speed operation causes bulky size and heavy weight. On the other hand, although the latter one takes the merits of reduced overall size and weight, the planetary gear inevitably involves transmission loss, acoustic noise, and regular lubrication. Magnetic gearing is becoming attractive since it offers the advantages of high efficiency, reduced acoustic noise, and maintenance free [53]. By artfully integrating the magnetic gear into a motor drive, the low-speed requirement for direct driving and the high-speed requirement for machine design can be achieved simultaneously [54]. Figure 25 shows the detailed configuration of a magnetic-geared in-wheel PM BLDC motor. The artfulness is the share of a common PM rotor, namely, the outer rotor of a PM BLDC motor and the inner-rotor of a concentrically arranged magnetic gear. The operating principle of this magnetic-geared PM BLDC drive is similar to that of a high-speed planetary-geared inner-rotor drive, but with the difference that this one is an outer-rotor drive. That is to say, the motoring operation is the same as the PM BLDC drives. First, the stator is fed by three-phase voltages, which are rated at 220 Hz, to achieve the rated speed of 4,400 r/min. Then, the magnetic gear steps down the rated speed to 600 r/min, which in turn boosts up the torque for direct driving. The torque transmission is based on the modulation of the air-gap flux density distributions along the radial and circumferential directions. The space harmonic is modulated by the 25 stationary steel pole pieces from three pole pairs in the inner air-gap to 22 pole pairs in the outer air-gap. Hence, the torque in the outer rotor can be significantly amplified to about seven times that of the inner-rotor.

ISG Systems

In conventional automobiles, the starter motor and generator are separately coupled with the ICE, hence providing high starting torque for cold cranking and



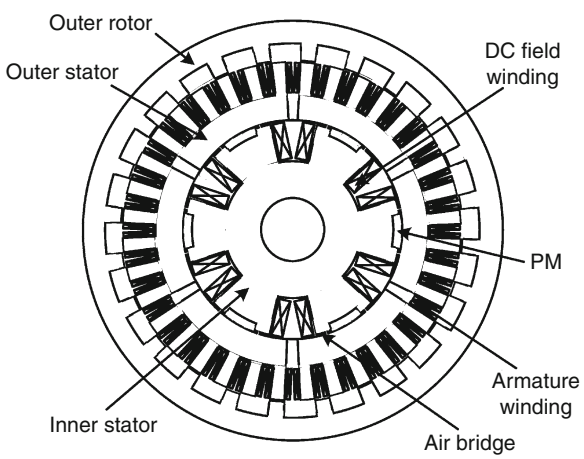
Vehicle Traction Motors. Figure 25
Magnetic-geared in-wheel motor drive. (**a**) Elevation view; (**b**) Side view

generating electricity for battery charging, respectively. This arrangement takes the advantage of simplicity but suffers from the poor utilization of both machines, hence resulting in heavy weight and bulky size. In order to incorporate both functions in a single unit, the development of integrated starter-generator (ISG) systems is accelerating.

By incorporating the inherent merits of PM brushless machines into the ISG, the resulting PM brushless ISG system is attractive for the latest micro- and mild HEVs. The stator doubly fed DSPM brushless machine [55] is a particular type of the aforementioned PM hybrid brushless machine topologies, which is promising for application to the ISG system. Its configuration is shown in Fig. 26, in which there are two magnetic field excitations, namely, the PMs and the DC field windings, air bridges in shunt with the PMs in the inner stator, AC armature windings in the salient-pole outer stator, and the salient-pole outer rotor with no PMs or windings.

This stator doubly fed DSPM brushless ISG system offers several distinct advantages:

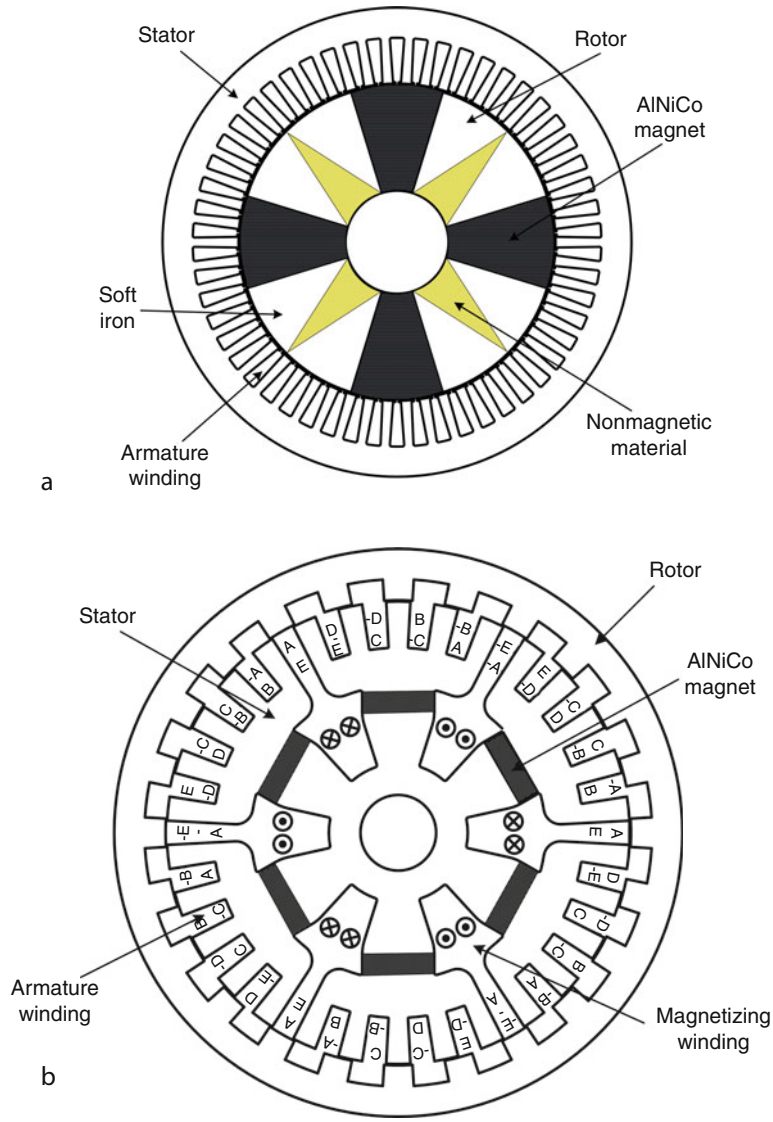
1. The DC field current can be bidirectionally controlled to strengthen and weaken the air-gap flux density, hence offering high starting torque for cold cranking and constant output voltage over a wide speed range for battery charging. Meanwhile, the air bridge amplifies the effect of flux weakening.
2. The outer-rotor topology can fully utilize the space of the inner stator to accommodate the PMs and the DC field windings, hence reducing the overall size of the machine.
3. Since the outer rotor does not involve any windings or PMs, it can provide high mechanical integrity which is essential to handle the high starting torque during cold cranking.
4. Since the stator adopts fractional-slot concentrated windings, it can effectively reduce the cogging torque which usually occurs in the PM BLDC machines. In addition, it can shorten the length of end windings, hence saving the copper material and improving the power density.



Vehicle Traction Motors. Figure 26
ISG system based on a stator doubly fed DSPM machine

Memory Brushless Machine Topologies

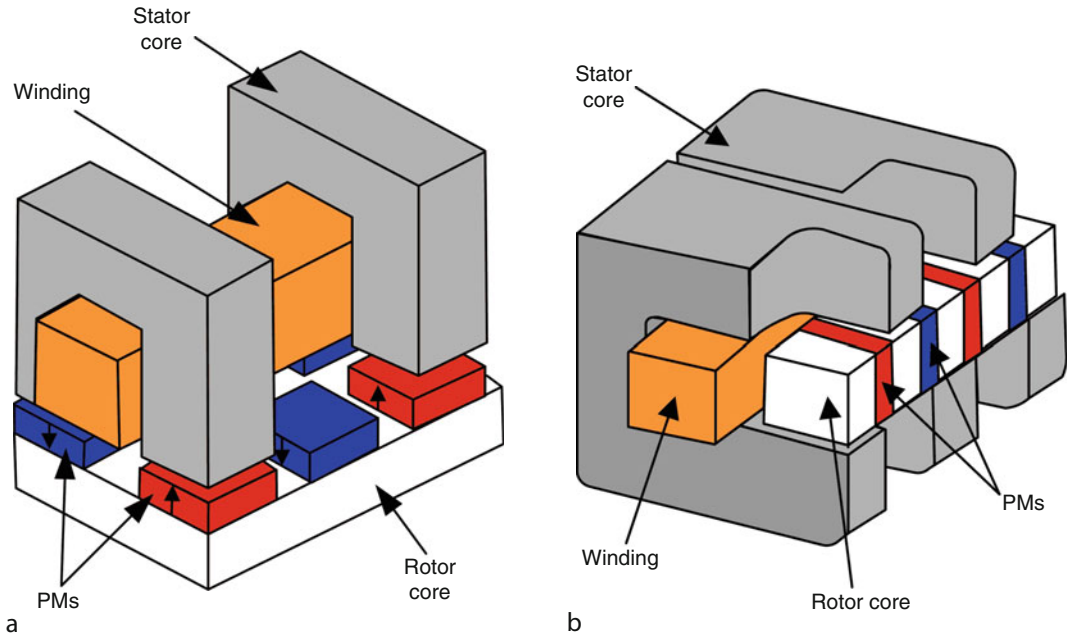
Although the DC field winding in PM hybrid motors enables the air-gap flux controllable, the use of DC field current inevitably causes additional power loss and degrades the efficiency. Hence, a new class of flux controllable PM machines, namely, the memory brushless machine was advent, which has the distinct ability to change the intensity of magnetization and also memorize the flux-density level in the PMs [56]. Figure 28 shows the topologies of memory brushless machines. In Fig. 27(a), it consists of aluminum nickel cobalt (AlNiCo) PMs sandwiched by soft iron, which are then mechanically fixed to a nonmagnetic shaft. The online magnetization is achieved by properly applying a short DC current pulse flowing through the stator armature winding to change the magnetization level of the AlNiCo PMs in the rotor. Figure 27(b) is a memory DSPM motor that adopts two-layer inner stator and outer rotor. In the stator, the armature windings are located in the outer layer, while both the PMs and magnetizing windings are placed in the inner layer, hence achieving a compact structure. Since the outer rotor is simply composed of salient poles without PMs or windings, it is very robust. The PM material used in the motor is an AlNiCo alloy [57]. Due to its direct magnetization of PMs by a temporary current pulse in the magnetizing windings, the flux control is highly effective and highly efficienct.



Vehicle Traction Motors. Figure 27
Memory brushless machines. (**a**) Rotor-PM; (**b**) Stator-PM

Transverse Flux Machine

PM transverse flux machine (TFM) [58, 59] is claimed to exhibit a higher torque density compared to the induction motor, SR motor, PM BLDC motor, and normal PM synchronous motor. In the TFM, the flux enters segments of stator core in transverse, i.e., perpendicular to the direction of rotor movement in the cross section plane, therefore, it is called transverse flux machine. The TFM stator iron core has two basic structures: (1) U-shape poles which have both teeth in the same axial plane and (2) claw-poles. Figure 28 shows the schematic configurations of the TFM. For each phase, a toroidal winding is placed inside the stator teeth or poles. Generally, TFMs have a relatively large number of poles, all of which interact with the total ampere-conductors of each phase. This enables very high electric loadings and, hence, high-torque densities to be achieved. Hence, it is especially suited



Vehicle Traction Motors. Figure 28
PM transverse flux motor. (**a**) Surface PMs; (**b**) Flux concentrated PMs

for in-wheel direct drive. However, they have a significant leakage flux and a relatively high winding inductance, as well as a poor power factor. This impacts significantly on the associated VA rating of the power electronics converter.

Bibliography

Primary Literature

1. Chan CC, Chau KT (2001) Modern electric vehicle technology. Oxford University Press, Oxford, UK
2. Zhu ZQ, Howe D (2007) Electrical machines and drives for electric, hybrid, and fuel cell vehicles. Proc IEEE 95:746–765
3. Chau KT, Chan CC, Liu C (2008) Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles. IEEE Trans Ind Electron 55:2246–2257
4. Fuhs A (2008) Hybrid vehicles and the future of personal transportation. CRC, Boca Raton
5. Xu W, Zhu J, Guo Y et al (2009) Survey on electrical machines in electrical vehicles. In: IEEE international conference on applied superconductivity and electromagnetic devices, Chengdu, 25–27 Sept 2009, pp 167–170
6. Yamada K, Watanabe K, Kodama T et al (1996) An efficiency maximizing induction motor drive system for transmission less electric vehicle. In: Proceeding the 13th international electric vehicle symposium, Osaka, vol II, pp 529–536
7. Jiang SZ, Chau KT, Chan CC (2002) Performance analysis of a new dual-inverter pole-changing induction motor drive for electric vehicles. Electr Power Compon Syst 30:11–29
8. Jiang SZ, Chau KT, Chan CC (2003) Spectral analysis of a new six-phase pole-changing induction motor drive for electric vehicles. IEEE Trans Ind Electron 50(1):123–131
9. Chan CC, Chau KT (1996) An advanced permanent magnet motor drive system for battery-powered electric vehicles. IEEE Trans Veh Technol 45:180–188
10. Chan CC, Chau KT, Jiang JZ et al (1996) Novel permanent magnet motor drives for electric vehicles. IEEE Trans Ind Electron 43:331–339
11. Chan CC, Jiang JZ, Xia W, Chau KT (1995) Novel wide range speed control of permanent magnet brushless motor drives. IEEE Trans Power Electron 10:539–546
12. Cheng M, Hua W, Zhang J, Zhao W (2011) Overview of stator-permanent magnet brushless machines. IEEE Trans Ind Electron. doi:10.1109/TIE.2011.2123853
13. Liao Y, Liang F, Lipo TA (1995) A novel permanent magnet machine with doubly salient structure. IEEE Trans Ind Appl 3(5):1069–1078
14. Cheng M, Chau KT, Chan CC (2001) Static characteristics of a new doubly salient permanent magnet motor. IEEE Trans Energy Convers 16(1):20–25
15. Deodhar RP, Andersson S, Boldea I, Miller TJE (1996) The flux-reversal machine: a new blushless doubly-salient permanent-magnet machine. In: Proceedings of the IEEE IAS Annual Conference. San Diego, 6–10 Oct 1996, pp 786–793

16. Kim TH, Jang KB, Chun YD et al (2005) Comparison of the characteristics of a flux reversal machine under the different driving methods. IEEE Trans Magn 41(5):1916–1919
17. Hoang E, Ben-Ahmed AH, Lucidarme J (1997) Switching flux permanent magnet polyphased machines. In: Proc. Eur. Conf. Power Electron. Appl., Trondheim, pp 903–908
18. Hua W, Cheng M, Zhu ZQ et al (2008) Analysis and optimization of back EMF waveform of a flux-switching permanent magnet motor. IEEE Trans Energy Convers 23(3):727–733
19. Zhu ZQ, Chen JT (2010) Advanced flux-switching permanent magnet brushless machines. IEEE Trans Magn 46(6): 1447–1453
20. Hua W, Zhu ZQ, Cheng M et al (2005) Comparison of flux-switching and doubly-salient permanent magnet brushless machines. In: Proc. Int. Conf. electrical machines and systems, Nanjing, pp 165–170
21. Zhu X, Cheng M (2010) Design, analysis and control of hybrid excited doubly salient stator-permanent-magnet motor. Sci China Tech Sci 53(1):188–199
22. Chan CC, Jiang Q, Zhou E (1995) A new method of dimension optimization of switched reluctance motors. In: Proceedings of Chinese international conference on electrical machines, Hangzhou, pp 1004–1009
23. Chan CC, Jiang Q, Zhan YJ, Chau KT (1996) A high-performance switched reluctance drive for P-star EV project. In: Proceedings of 13th international electric vehicle symposium, Osaka, vol II, pp 78–83
24. Zhan YJ, Chan CC, Chau KT (1999) A novel sliding-mode observer for indirect position sensing of switched reluctance motor drives. IEEE Trans Ind Electron 46:390–397
25. Krishnan R (1996) Review of flux-weakening in high performance vector controlled induction motor drives. In: Proc. IEEE Int. Symp. Industrial Electronics, Warsaw, pp 917–922
26. Miller JM, Gale AR, McCleer PJ et al (1998) Starter/alternator for hybrid electric vehicle: comparison of induction and variable reluctance machines and drives. In: Proceedings of the industry applications Society Annual Meeting, Oct 1998, St Louis, pp 513–523
27. Zhu X, Cheng M, Zhao W et al (2007) A transient co-simulation approach to performance analysis of hybrid excited doubly salient machine considering indirect field-circuit coupling. IEEE Trans Magn 43(6):2558–2560
28. Zhao W, Cheng M, Zhu X et al (2008) Analysis of fault tolerant performance of a doubly salient permanent magnet motor drive using transient co-simulation method. IEEE Trans Ind Electron 55(4):1739–1748
29. Bose BK (1992) Modern power electronics: evolution, technology, and applications. IEEE, New York
30. Chan CC, Chau KT, Chan DTW, Yao JM (1997) Soft switching inverters in electric vehicle. In: Proceedings of the 14th international electric vehicle symposium, CD-ROM, Orlando
31. Lai JS (1997) Resonant snubber-based soft-switching inverters for electric propulsion drives. IEEE Trans Ind Electron 44:71–80
32. Murai Y, Cheng J, Yoshida MA (1997) Soft-switched reluctance motor drives circuit with improved performances. In: Proc of IEEE power electronics specialists conference, 22–27 June 1997, St Louis, pp 881–886
33. Chau KT, Ching TW, Chan CC, Chan DTW (1997) A novel two-quadrant zero-voltage transition converter for DC motor drives. In: Proceedings of IEEE international conference on industrial electronics, New Orleans, pp 517–522
34. Divan DM (1986) The resonant DC link converter – a new concept in static power conversion. In: Proceedings of IEEE industry application society annual meeting, Denver, pp 648–656
35. Cho JG, Kim WH, Rim GH, Cho KY (1997) Novel zero transition PWM converter for switched reluctance motor drives. In: Proceedings of IEEE power electronics specialists conference, St Louis, pp 887–891
36. Ching TW, Chau KT, Chan CC (1998) A new zero-voltage-transition converter for switched reluctance motor drives. In: Proceedings of IEEE power electronics specialists conference, Fukuoka, pp 1295–1301
37. Rahman MF, Haque ME, Tang L, Zhong L (2004) Problems associated with the direct torque control of an interior permanent-magnet synchronous motor drive and their remedies. IEEE Trans Ind Electron 51(4):799–809
38. Pascas M, Weber J (2005) Predictive direct torque control for the PMsynchronous machine. IEEE Trans Ind Electron 52(5):1350–1356
39. Cavallaro C, Tommaso AOD, Miceli R et al (2005) Efficiency enhancement of permanent-magnet synchronous motor drives by online loss minimization approaches. IEEE Trans Ind Electron 52(4):1153–1160
40. Shu Y, Cheng M, Kong X (2008) Online efficiency optimization of stator-doubly-fed doubly salient motor based on a loss model. In: Proceedings of 11th international conference on electrical machines and systems, Wuhan, pp 1174–1178
41. Cheng M, Sun Q, Zhou E (2006) New self-tuning fuzzy PI control of a novel doubly salient permanent-magnet motor drive. IEEE Trans Ind Electron 53(3):814–821
42. Pajchrowski T, Zawirski K (2007) Application of artificial neural network to robust speed control of servo drive. IEEE Trans Ind Electron 54(1):200–207
43. Acarnley PP, Watson JF (2006) Review of position-sensorless operation of brushless permanent-magnet machines. IEEE Trans Ind Electron 53(2):352–362
44. Silva C, Asher GM, Sumner M (2006) Hybrid rotor position observer for wide speed-range sensorless PM motor drives including zero speed. IEEE Trans Ind Electron 53(2): 373–378
45. Angelo CD, Bossio G, Solsona J et al (2006) Mechanical sensorless speed control of permanent-magnet AC motors driving an unknown load. IEEE Trans Ind Electron 53(2):406–414
46. Emadi A, Young Joo L, Rajashekara K (2008) Power electronics and motor drives in electric, hybrid electric, and plug-in hybrid electric vehicles. IEEE Trans Ind Electron 55(6):2237–2245
47. U.S. Department of Energy (2007) Plug-in hybrid electric vehicle R&D plan. http://www1.eere.energy.gov/vehiclesandfuels/pdfs/program/phev_rd_plan_june_2007.pdf

48. Mecrow BC, Jack AG, Haylock JA, Coles J (1996) Fault-tolerant permanent magnet machine drives. IEEE Proc Electric Power Appl 143(6):437–442
49. Akita H, Nakahara Y, Miyake N, Oikawa T (2003) New core structure and manufacturing method for high efficiency of permanent magnet motors. In: Conference record of IEEE IAS annual meeting, Amagasaki, 12–16 Oct 2003, pp 367–372
50. Miller JM (2006) Hybrid electric vehicle propulsion system architectures of the e-CVT type. IEEE Trans Power Electron 21(3):756–767
51. Xu L, Zhang Y, Wen X (2007) Multi-operational modes and control strategies of dual mechanical port machine for hybrid electrical vehicles. In: Proceedings of the IEEE IAS Annual meeting, New Orleans, pp 1710–1717
52. Cheng Y, Cui S, Song L, Chan CC (2007) The study of the operation modes and control strategies of an advanced electromechanical converter for automobiles. IEEE Trans Magn 43(1):430–433
53. Atallah K, Howe D (2001) A novel high performance magnetic gear. IEEE Trans Magn 37(4):2844–2846
54. Chau KT, Zhang D, Jiang JZ, Liu C, Zhang Y (2007) Design of a magnetic-geared outer-rotor permanent-magnet brushless motor for electric vehicles. IEEE Trans Magn 43(6):2504–2506
55. Chau KT, Li YB, Jiang JZ, Liu C (2006) Design and analysis of a stator doubly fed doubly salient permanent magnet machine for automotive engines. IEEE Trans Magn 42(10):3470–3472
56. Ostovic V (2003) Memory motor. IEEE Ind Appl Mag 9(1):52–61
57. Yu C, Chau KT, Liu X et al (2008) A flux-mnemonic permanent magnet brushless motor for electric vehicles. J Appl Phys 103(07103):1–3
58. Henneberger G, Bork M (1997) Development of a new transverse flux motor. In: IEE Colloq. new topologies for permanent magnet machines, London, pp 1/1–1/6
59. Baserrah S, Orlik B (2009) Comparison study of permanent magnet transverse flux motors (PMTFMs) for in-wheel applications. In: Proceedings of the international conference on power electronics and drive systems, Taipei, pp 96–101

Books and Reviews

Chan CC (2002) The state of the art of electric and hybrid vehicles. Proc IEEE 90:247–275
Chan CC, Chau KT (1997) An overview of power electronics in electric vehicles. IEEE Trans Ind Electron 44:3–13
Chau KT, Ming C (2010) New drive technology for electric vehicles. China Machine, Beijing (In Chinese)
Ehsani M, Rahman KM, Toliyat HA (1997) Propulsion system design of electric and hybrid vehicles. IEEE Trans Ind Electron 44:19–27
Ehsani M, Gao Y, Gay SE, Emadi A (2005) Modern electric, hybrid electric, and fuel cell vehicles: fundamentals, theory, and design. CRC, Boca Raton
Husain I (2003) Electric and hybrid vehicles-deign fundamentals. CRC, Boca Raton
Rashid MH (2005) Modern electric, hybrid electric, and fuel cell vehicles: fundamentals, theory, and design. CRC, Boca Raton

Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in

Ping Wang, Fuqiang Liu, Wenjie Shu, Wenqiang Xu
Department of Information and Communication Engineering, Tongji University, Shanghai, China

Article Outline

Glossary

BGF Beacon-Assist Geographic Forwarding, one of the approaches for enhanced GPSR in packet-forwarding strategy to enable nodes to forward packets by considering their neighbors' beacon interval in neighbor list.

Geographical routing A kind of routing protocol in VANET that uses geographic information to help complete packets delivery. This kind generally contains two parts: location server and packets forwarding.

GPSR The famous geographical routing protocol, named greedy perimeter stateless routing, uses greedy forwarding strategy to select a next hop and perimeter forwarding strategy to cope with routing failure caused by local minimum.

ITR Inaccuracy-Tolerant Range represents the inaccurate degree of location information.

Local minimum No neighbor exists which is closer to the destination than the intermediate node itself.

NUR Neighbor-Awareness position Update, one of the approaches for enhanced GPSR in beaconing to set the position update interval dynamically.

SDR The Successful Date Rate defines as the ratio of packets successfully delivered to the destination node to those generated by the sender.

VANET The VehiculAr NETwork is a special communication network comprising of radio-enabled vehicles and roadside units, including vehicle-to-vehicle communication and vehicle-to-roadside unit communication.

Definition of the Subject

In vehicular network (VANET), moving vehicles always have a high speed and a frequent change of their moving direction. These make the network topology highly dynamic. Because of the highly dynamic nature of the mobile nodes, finding and maintaining routes is very challenging in VANET.

GPSR is considered as a promising geographical routing protocol applicable for VANET. In order to cope with the problem of highly dynamic topology, it makes routing decisions at the intermediate nodes instead of building a constant route. Under this strategy, routing failure caused by nodes disconnection in constant routing approach can be efficiently mitigated. However, there are still some problems in GPSR. As in greedy forwarding strategy, the intermediate node always selects the next hop node that lies close to the relaying nodes' transmission range border; the selected one has high possibility to leave the transmission range because of the high speed node movement. GPSR does not take this into consideration. Another problem comes from the constant beacon interval strategy. Such principle may lead to high routing overhead and poor performance in various nodes density environment.

Introduction

Researches on routing protocol in VANET come from the bases of earlier researches on routing protocol in Mobile Ad hoc network (MANET). They share the characteristics of self-organization, self-management, low-bandwidth, and short radio transmission range. However, VANET differs from MANET by its highly dynamic topology. The traditional topology-based routing strategies in MANET, such as AODV (Ad hoc On-demand Distance Vector) and DSR (Dynamic Source Routing), are not approving enough to be used in VANET. These routing protocols are under the principle of building a constant route table and maintaining the pre-build route table. Due to the highly dynamic mobility, routing protocols under such principle are unable to quickly find, maintain, and update long routes in VANET. The route will become invalid if some intermediate nodes disconnected. Thus, frequent route recovery will make the whole performance inefficient. Take AODV for example. It has been tested in real-world experiments. Results show that packets are excessively lost due to route failures.

Vehicles are not only simply mobile nodes, but also they have characteristics of high speed, limited movement along roads, and capability to equip with GPS system. As GPS systems are widely used in vehicles now, the mobile nodes can easily get their position information. This trend brings a new type of routing protocol, geographical routing, or position-based routing, which uses the position information to help deliver packets. Position-based routings mainly content two parts: location server and forwarding strategy. Location server is responsible for providing location information of the destination nodes to the source nodes. After getting location information of the destination nodes, different kinds of forwarding strategies will be used at the intermediate nodes to make the forwarding decisions in order to select an appropriate next hop. As introduced above, one of the main characteristics of position-based routings is that they do not try to build a constant route table. In this way, it makes position-based routings more applicable to VANET with highly dynamic mobility. Researchers see position-based routing as a more promising approach and do researches on them.

GPSR is one of the most typical position-based routing protocols. It uses the positions of routers and a packet's destination to make packet forwarding decisions. It contains two modes of forwarding strategy: greedy forwarding mode and perimeter forwarding mode. GPSR makes greedy forwarding decisions using only information about a router's immediate neighbors in the network topology. When a packet reaches a region where greedy forwarding is impossible, the algorithm recovers by routing around the perimeter of the region. By keeping state only about the local topology, GPSR scales better in per-router state than shortest-path and topology-based routing protocols as the number of

network destinations increases. Under mobility's frequent topology changes, GPSR can use local topology information to find correct new routes quickly.

Although GPSR shows a much higher performance than the traditional topology-based routing protocols in VANET, it still has some drawbacks within its routing strategy. GPSR uses a fixed period beaconing mechanism to send HELLO messages. Such mechanisms may lead to several problems such as wasted bandwidth, delaying of data packet, and increased network congestion. As GPSR adopts greedy forwarding algorithm, in which the intermediate node always selects the next hop node that lies close to the relaying nodes' transmission range border. The selected route may be not stable enough for the chosen next hop nodes have a high possibility to move away from the relaying nodes. One more drawback comes from the perimeter forwarding strategy. GPSR lacks considering the usage of road topology that restricts nodes' movements. Under such perimeter approach, the select route path may be not the most efficient and shortest one.

Describing Enhanced GPSR Using Neighbor-Awareness Position Update and Beacon-Assist Geographic Forwarding

The enhanced GPSR is modified by the algorithm called GPSR-N&B. It includes two parts: Neighbor-Awareness position Update (NAU) and Beacon-Assist Geographic Forwarding (BGF). GPSR-N&B uses NAU to set the position update interval dynamically, according to the neighbors' number and position of the relaying nodes. And BGF strategy enables nodes to forward packets by considering their neighbors' beacon interval in neighbor list. This algorithm aims to reduce beacon overhead and medium access control (MAC) layer collisions in different densities and velocities of VANET, while ensuring the Successful Date Rate (SDR) performances well.

Neighbor-Awareness Position Update (NAU)

GPSR-N&B is based on the following assumptions (a) all links are bidirectional and (b) all nodes can easily get their current location information and velocity.

In VANET, a vehicle's movement has effect on its neighbors' movement. NAU adapts the beacon update intervals according to the number, position, and velocity of the nodes in the neighborhood. The Neighbor-Awareness position Update (T_i) is shown as follows:

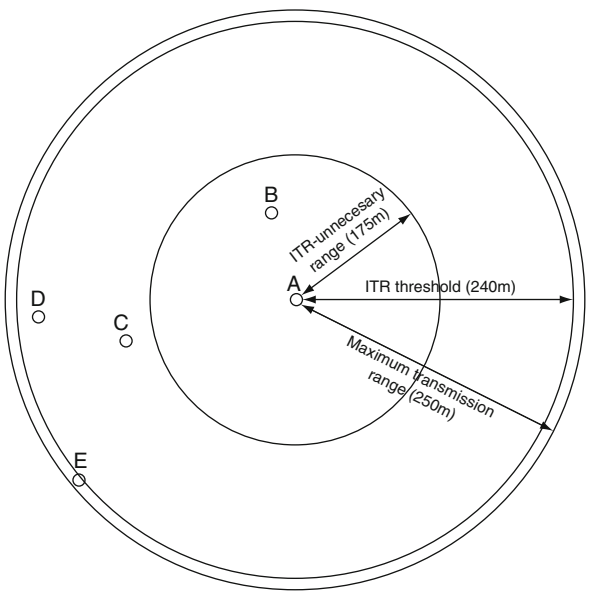
$$T_i = \frac{1}{2} T_i' + \frac{1}{2} T_i''. \quad (1)$$

T_i' is affected by the neighbors' position, while T_i'' is affected by the number of neighbors.

1. T_i'

In GPSR-N&G, an aspect of inaccuracy-tolerant range (ITR) is supposed to represent the inaccurate degree of location information that can be accepted. Intuitively, within maximum transmission range, the larger the distance between two nodes is and the more the probability the link may break, the shorter the beacon interval should be.

Figure 1 demonstrates the inaccuracy-tolerant range around node *A*. For example, node *E* which is located between the maximum transmission range (250 m) and ITR threshold (240 m) will not forward packets from node *A*, because of the changeable environment and distance between them, which may often lead to link breakdown.



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 1 Inaccuracy-tolerant range

As node D is within the range of 200–240 m, ITR equals to 240-d, where d is the distance between nodes A and D: $d = \sqrt{(x_A - x_D)^2 + (y_A - y_D)^2}$. The relative beacon interval time T_{AD} will be set to be $\text{Max}\left[\frac{(240-d)\,T_{\text{MINbeacon}}}{240-175}, T_{\text{MINbeacon}}\right]$.

As node B is within the range of 175 m, the default beacon interval $T_{\text{MINbeacon}}$ is then used as T_{AB}, instead of calculating the effect of ITR.

In Fig. 2, node A and B are neighbors. After considering A and B, a relative beacon interval T_{AB} will be got as well as T_{AC} and T_{AD} in a similar way. Then T'_A can be set as:

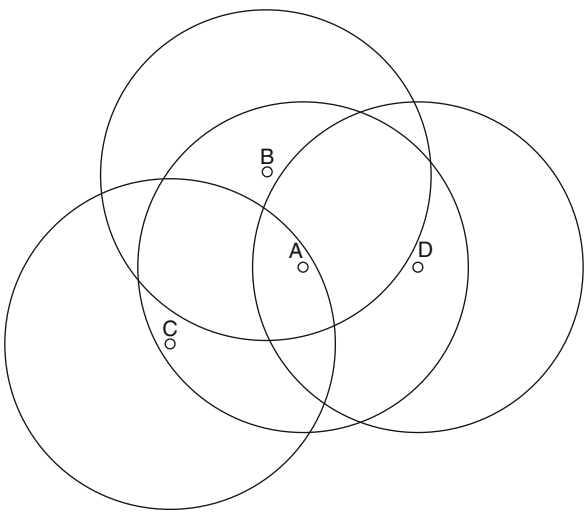
$$T'_A = \frac{T_{AB} + T_{AC} + T_{AD}}{3}. \tag{2}$$

Then, $T'_i = \frac{\sum_{j=2}^{k} T_{ij}}{K}$, where K is the number of node i's neighbors.

2. T''_i

T''_i is set as $T''_i = wK$, in which w is computed in simulation test. Finally, node i's new beacon interval is shown in Eq. 3:

$$T_i = \frac{1}{2}\frac{\sum_{j=2}^{k} T_{ij}}{K} + \frac{1}{2}wK \tag{3}$$



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 2 Neighbor-awareness position update

Upon initialization, each node broadcasts a beacon informing its neighbors about its current location (x_i, y_i) and velocity (v_{xi}, v_{yi}) using default beacon interval $T_{\text{DEFbeacon}}$, which is set to $\frac{175}{V_{\text{MAX}}}$ in the following simulation. Here, V_{MAX} is the maximum node velocity.

After initialization, each node replaces its own beacon interval by T_i.

Beacon-Assist Geographic Forwarding (BGF)

In NAU, the way to set position update time is based on the change of relative distance, instead of change of position. The neighbor list is changed by using relative distance in x and y directions, instead of the position as Table 1 shows.

In Table 1, X-Distance and Y-Distance are the distance between current node and its neighbor in x direction and y direction. V_x and V_y are the relative velocity of current node and its neighbor in x direction and y direction. T_{beacon} is the neighbor's beacon interval time.

As in NAU, each neighbor's T_{beacon} includes the topology information in its vicinity. So the neighbors' T_{beacon} can be used as the topology information of the next hop in choosing the efficient forwarding node. In BGF, a routing metric is defined which depends on (a) relative distance and (b) T_{beacon} as follows:

$$M_j(l_j, T_{j\text{beacon}}) = \alpha g(l_j) + \beta h(T_{j\text{beacon}}) \tag{4}$$

$$g(l_j) = \exp(-(l_j - l_i)) \tag{5}$$

$$h(T_{j\text{beacon}}) = \exp\left(-(T_{j\text{beacon}} - T_{\text{DEFbeacon}})\right) \tag{6}$$

Here α and β are the weights to distance and T_{beacon}. l_i is the distance between the destination and node i. l_j is

Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Table 1 Node A's neighbor list

	X-distance	Y-distance	V_x	V_y	*Tbeacon*
B	−50	210	10	15	2
C	−202	−45	−20	5	3
–	–	–	–	–	–

the distance between the destination and node j in node i's neighborhood, and $T_{j\text{beacon}}$ is node j's beacon interval.

Simulation Results

Performance of GPSR-N&B is evaluated by simulation varying the beacon interval, the maximum speed of mobile nodes and the network density for freeway models. NS-2 is used to estimate the effect of inaccurate location information with different beacon intervals, speeds, and densities on GPSR, which has been proven to perform efficiently with accurate location information.

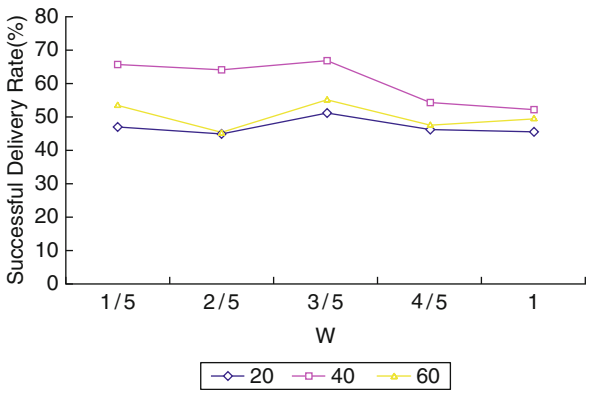
1. Set w in T_i''
   In the simulation, w of 1/5,2/5,3/5,4/5,5/5 is tested in different number of nodes (20,40,60). In Fig. 3, w equaling to 3/5 outperforms the others in terms of the SDR. As a result, in following simulation, w is set to 3/5.
2. Results and Analysis
   The results represented here are averaged over ten runs, each using a different random seed. Nodes are randomly placed over a 1,000*50 m^2 freeway created by sumo and move.

Performance of various beacon intervals (0.5, 1, 2, 4, 8) is compared as well as GPSR-N&B by varying maximum moving speed (10, 20, 30 m/s) and different number of nodes (20, 40, 60). Each simulation lasts 100 simulation seconds. The size of data packets is 32 bytes. Five random source–destination pairs, using CBR as the traffic flows and lasting about 20 s, are selected. Different aspects of the performance of the routing protocol are evaluated using the following metrics:

Successful date rate (SDR): The ratio of packets successfully delivered to the destination node to those generated by the sender.

Routing overhead: The total number of beacon packets transmitted.

Number of MAC layer collisions: The number of collisions in MAC layer due to the beacon packets.



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 3
SDR of different w

Figure 4a, b, and c illustrates that GPSR-N&B outperforms from GPSR with fixed beacon interval with various numbers of nodes and maximum node speed. GPSR with fixed beacon interval results in low SDR, and cannot cope with the topology change with high mobility.

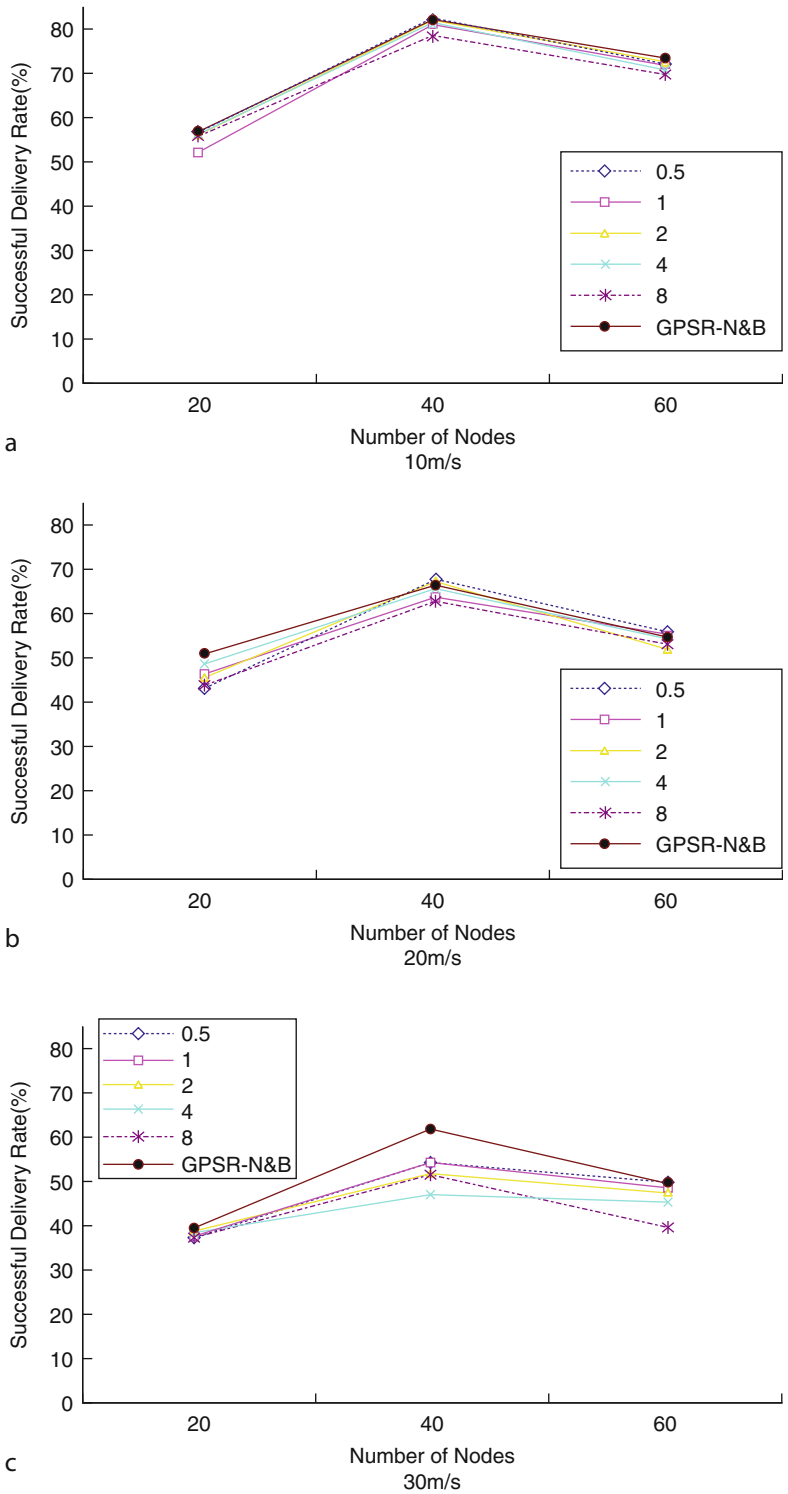
Firstly PSR-N&B controls the routing overhead and decreases the MAC layer collisions to leave more channel resource for the CBR packets which will be discussed later. Secondly, GPSR-N&B using BGF makes use of the beacon intervals of neighbors for estimating the next hop topology, which makes the greedy forwarding more stable and efficient. The greedy forwarding is extended to two-hop area.

Figure 5a, b, and c shows that the routing overhead of GPSR-N&B is very low in all the simulations. The most important thing is that routing overhead of GPSR-N&B does not increase a lot with the increasing density of nodes, so GPSR-N&B fits the large number of nodes. In NAU, the node beacon interval is set according to the number of neighbors.

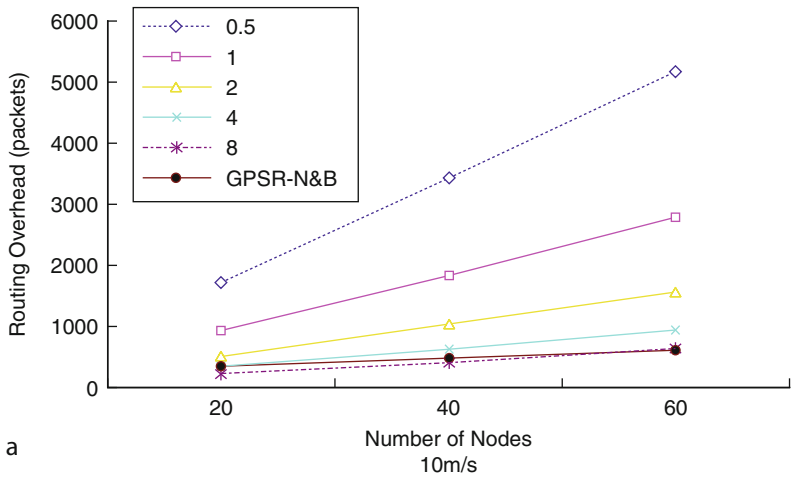
GPSR-N&B makes lower MAC layer collisions as shown in Fig. 6a, b, and c, due to the different beacon intervals of nodes in the network compared with GPSR using fixed beacon intervals. The overhead packets in high-density neighbor area are less than in low-density neighbor area due to the addition of T″ in NAU.

Survey of Different Strategies to Enhance GPSR

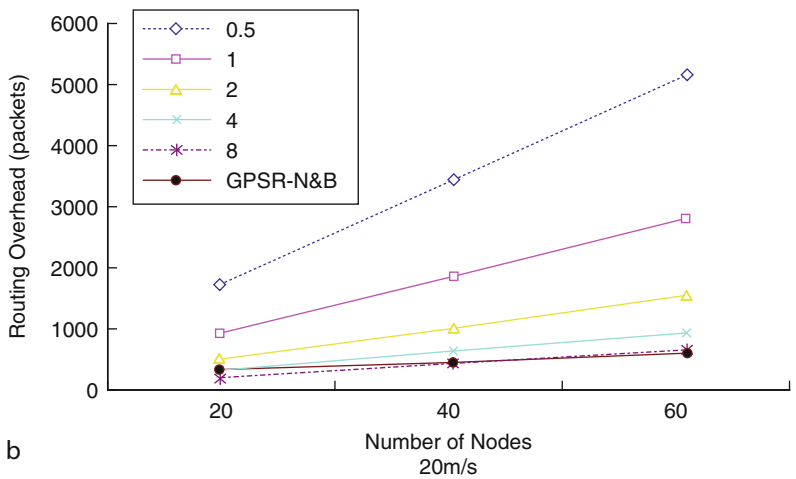
GPSR does not take cars' mobility characteristics into consideration which limits the performance in VANET. The movement of nodes is assumed to be arbitrary in MANET while bound to a street in VANET. It makes sense that many researchers are going to exploit the different characteristics to increase packet-forwarding



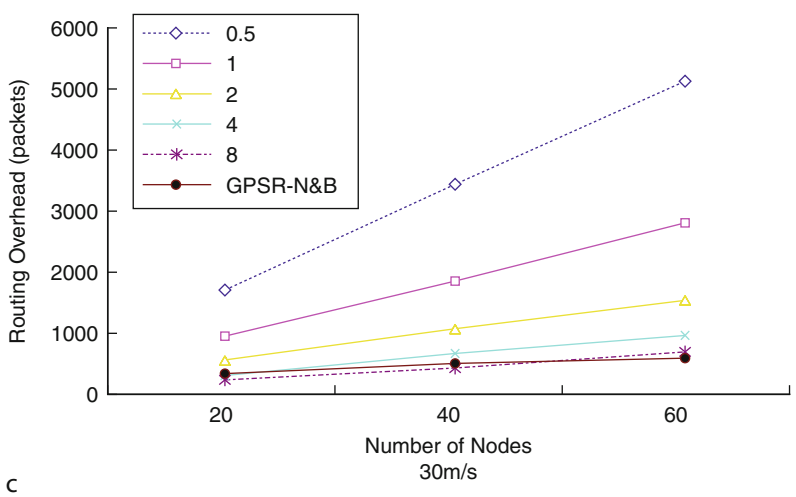
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 4
SDR varying maximum node speed



a

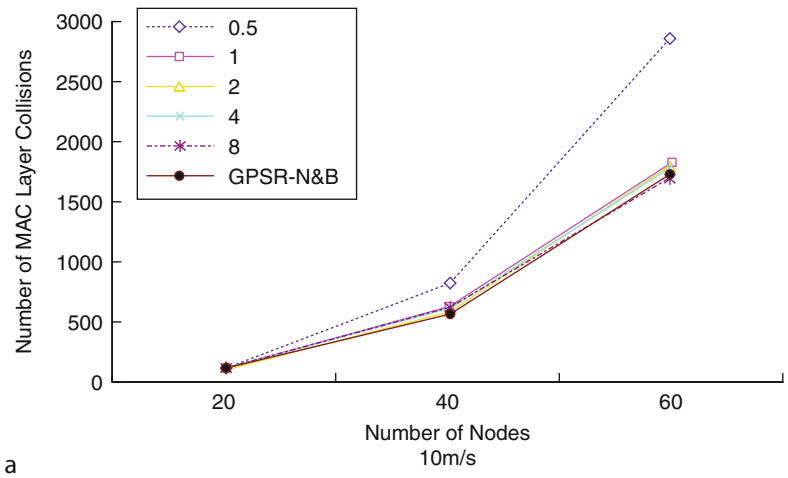


b

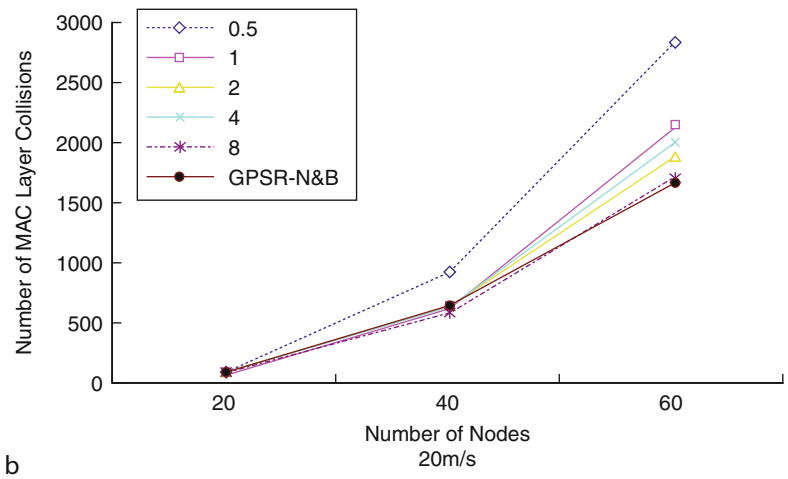


c

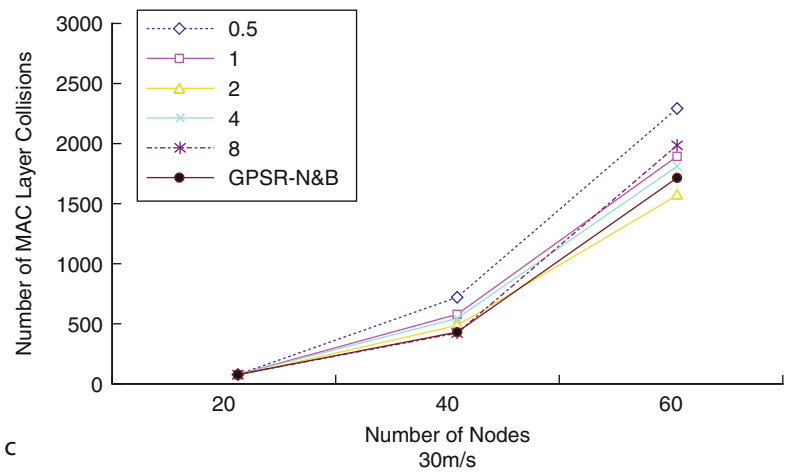
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 5
Routing overhead varying maximum node speed



a



b



c

Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 6
Number of MAC layer collisions varying maximum node speed

performance. Going from simple to complex, scenarios of cars movement can be classified as highway and city scenario.

Considering these characteristics in highway and city scenario, which will be discussed in following sections, researchers have developed other approaches to improve the routing performance in VANET compared with GPSR.

Strategies of Enhancement in Highway Scenario

Definition of the Subject In highway scenario, cars may move in different directions or in different lanes, but the main extent of movement is along the road, not across, which makes the movement one dimensional. This makes packet forwarding fairly easy but much sensitive to speed because cars may move in highway very fast.

Advanced Greedy Forwarding (AGF)

V. Naumov, R. Baumann, and T. Gross in [1] proposed the Advanced Greedy Forwarding (AGF) algorithm to significantly improve GPSR performance in VANETs.

Both the source and the destination nodes inform each other with the help of the location discovery service (e.g., reactive location service [2]) about their moving directions and speeds – velocity vectors. Velocity vector information is also added into HELLO beacons of all nodes. Velocity vector requires two additional bytes to store the information about nodes' speed and direction. The first byte encodes the direction in the range of 0–127. The second byte stores the speed in km/h (enough for representing the allowed maximum speed in most countries).

Also the information about the packet travel time is added in a data packet header. Every node forwarding a data packet adds its own processing time into packet header. A next hop node is chosen based on the velocity vectors information stored in the neighbor tables.

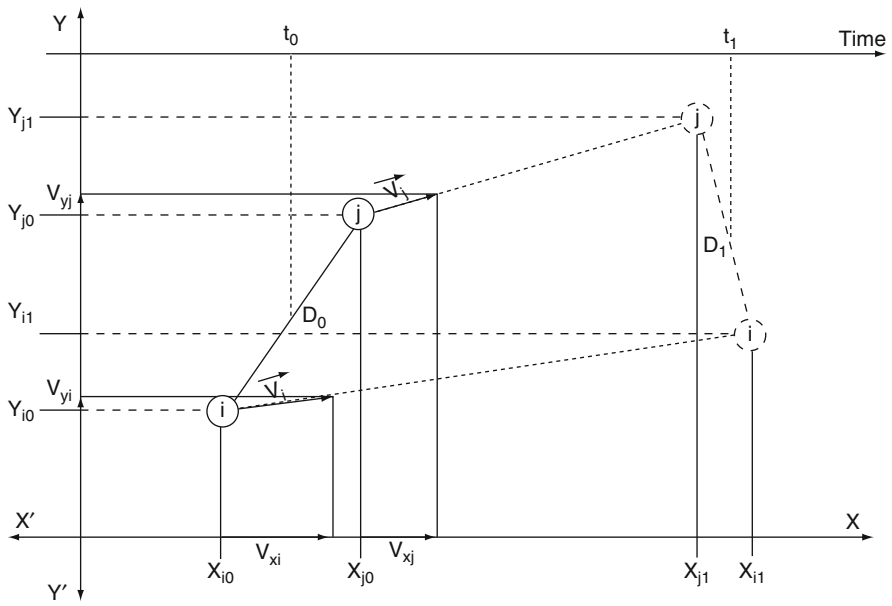
A node receiving a data packet checks if the destination is listed in its neighbor table and the entry is still valid, taking into account the packet travel time and the node's and the destination's velocity vectors. If this is the case, the node sends the packet directly to the destination. If the destination is in the neighbor table, but the new position estimation tells that the destination is most likely already out of the range, then the node closest to the new position of the destination is chosen as a next hop.

If no destination is found in the neighbor table, the node consults the packet travel time and estimates whether it may potentially reach the position of the destination recorded in the packet header within one hop, taking into account a distance potentially traveled by the destination node within the packet travel time. If yes, the non-propagating broadcast is sent around, with the search for the destination. If no answer is received (either from the destination or the node that has the destination in its table and is closer to the destination than the current node), then the next closest to the destination node is chosen, and the process repeats.

Movement Prediction–Based Routing

This routing concept, based on vehicles movement prediction, estimates the stability of each communication link in the network in terms of communication lifetime, and then selects the most stable route composed by the most stable intermediate links from the source till the destination. This concept can be applied to position-based routing protocols and has been implemented with GPSR under the network simulator NS2. By this way, GPSR's performance is improved.

Movement Prediction-Based Routing (MOPR) [3] determines the most stable path from a source to a destination in terms of communication lifetime by selecting the most stable intermediate links, then, the best intermediate vehicles. For example, assuming there is a network protocol which is capable to provide several unicast paths to a destination, one of those paths can result to be more stable with respect to the others. A stable path can increase the probability that link failures will be avoided during the whole communication. MOPR, based on vehicles' movement information, guarantees the selection of the best next hop for data forwarding. Using MOPR, each vehicle estimates the Link Stability (LS) for each neighboring vehicle before selecting the next hop for the data forwarding/sending. The LS is a relation between the link communication lifetime and a constant value (σ) which represents in general cases the routing route validity time, and it depends on the used routing protocol. Figure 7 shows how link lifetimes are estimated based on neighbors' movement information.



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 7
Link lifetime estimation

The lifetime of the link (i, j) (LifeTime$[i,j]$) corresponds to the estimated time $\Delta t = t_1 - t_0$ with t_1 is the time when D_1 becomes equal or bigger than the communication range R (i.e., the time when j goes out of the communication rage of i). D_1 and Δt are estimated using the initial positions of i and j ((X_{j0}, Y_{j0}), and their initial speeds $\bar{V}_j$ and $\bar{V}_j$, respectively):

$$D_1^2 = ((X_{i0} + Vx_i\Delta t) - (X_{j0} + Vx_j\Delta t))^2 + ((Y_{i0} + Vy_i\Delta t) - (Y_{j0} + Vy_j\Delta t))^2$$

$$D_1^2 = A\Delta t^2 + B\Delta t + C$$

$$\text{with} \begin{cases} A = (Vx_i - Vx_j)^2 + (Vy_i - Vy_j)^2 \\ B = 2[(X_{i0} - Y_{j0})(Vx_i - Vx_i) \\ \qquad + (X_{i0} - Y_{j0})(Vy_i - Vy_i)] \\ C = (X_{i0} - X_{j0})^2 + (Y_{i0} - Y_{j0})^2 \end{cases}$$

By solving the equation $A\Delta t^2 + B\Delta t + C - R^2 = 0$ it is easy to get Δt which corresponds to the LifeTime$[i,j]$. LS is calculated as follows:

$$LS[i,j] = \frac{\text{Life Time}[i,j]}{\sigma}, \text{ with } LS[i,j] = 1 \text{ when Life Time } [i,j] \geq \sigma$$

Once LS is calculated for each neighboring vehicle, MOPR selects as a next hop for data forwarding/sending the one corresponding to the highest LS (corresponding to the most stable neighboring link).

This approach should help as well in minimizing the risk of broken links and in reducing data loss and link-layer and transport retransmissions.

F. Granelli et al. have proposed a Movement-Based Routing Algorithm (MORA) [4] for vehicular ad hoc networks. It is applied to GPSR. MORA takes into account the physical location of neighboring vehicles and their movement direction when selecting the next hop for sending/forwarding packets. MOPR believes that considering only the position and the movement direction is not enough for a best next hop selection in VANETs. The vehicle's driving speed is important and should be taken into account as well. A vehicle which is almost out of communication range should not be selected as a next hop, which cannot be guaranteed without taking into account the speed. Thus, with MOPR, a vehicle which is estimated to go out of communication range in a short duration will not be selected as a next hop for data routing if a better candidate is available.

To show the performance improvements of MOPR over position-based routing protocols, it is applied over GPSR. It is not suitable to apply MOPR to GPSR as it is done to unicast routing [3, 5], where MOPR tries to select the path with the longest lifetime. When applying MOPR to GPSR as it is, the selected paths should be same or longer in terms of number of hops when compared to basic GPSR. And the calculation of neighboring links' LS before sending/forwarding each packet takes a considerable time. All this decreases the routing performances. To face this problem, MOPR is applied in a different way. When a vehicle wants to send or forward data, it first estimates the future geographic location after a duration time T in seconds for each neighbor. T is counted in seconds, and it is fixed to 1 s in the simulations. Then, it selects as next hop the closest neighbor to the destination which does not have a future location out of its communication rage after the time T.

By doing this, MOPR-GPSR avoids the case when a next hop goes out of communication range during a data packet transmission, thus, decreasing data loss and link-layer and transport retransmissions, which increases the routing performances.

To evaluate the performances of MOPR over position-based routing protocols, it is implemented on top of GPSR, named MOPR-GPSR. A Hierarchical Location Service (HLS) is used to provide the exact position information of the neighboring and the destination vehicles. More information on HLS is given in [6].

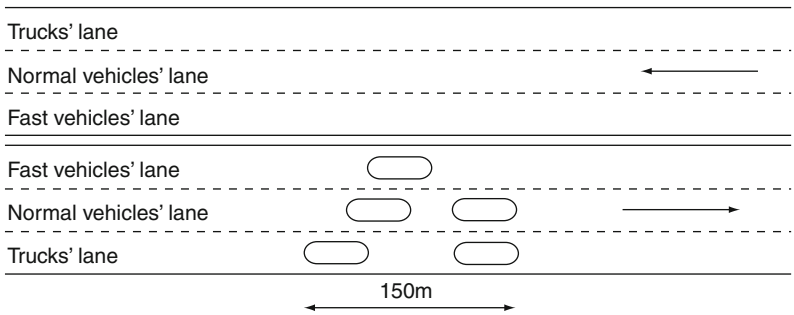
The simulations use a 5,000 m length highway scenario, with 200 vehicles moving on it as shown in Fig. 8. In each direction, there are three lanes with different speed ranges starting from a minimum speed value of 70 km/h and a maximum speed value which is increased from 120 to 220 km/h. In each direction there is a density of five vehicles every 150 m. The classical 802.11 Medium Access Control (MAC) functionalities are used. Traffic type was CBR with 1,024 Bytes of packet size and a 512 bps of maximum CBR rate. A transmitting source and a destination vehicle are selected randomly along the middle lane (normal vehicles' lane) in each direction.

Figure 9 shows the Packet Delivery Ratio (PDR) obtained for each routing protocol as function of vehicles' maximum speed. It is clearly shown that both MOPR and MORA guarantee a better PDR when compared to basic GPSR. As shown, higher the vehicles' maximum speed, higher the PDR of MOPR when compared to MORA. This means that MOPR guarantees the best PDR when speed is higher.

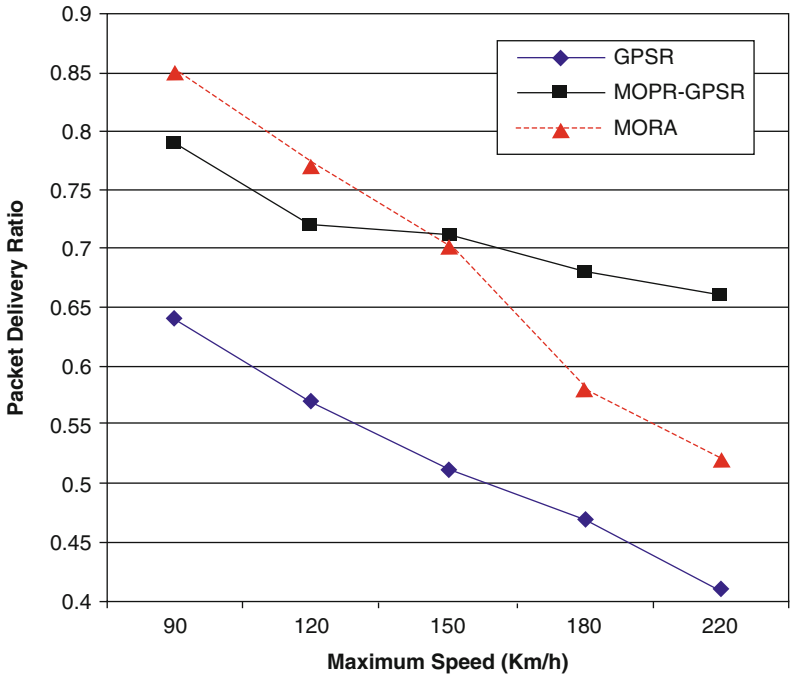
Figure 10 shows the delay for each routing protocol as a function of vehicles' maximum speed. MOPR improves the delay by at least two times when compared to both basic GPSR and MORA.

Figure 11 shows the routing overhead as a function of vehicles' maximum speed. MOPR decreases the routing overhead when compared to basic GPSR, but MORA decreases the routing overhead more. That means that MORA is the best in terms of routing overhead. But, in Fig. 12 the Hierarchical Location Service (HLS) overhead caused in the network which should be taken into account to evaluate the real performance of this routing protocol in terms of routing overhead. And it is clearly shown that MOPR is the best in terms of HLS overhead when compared to both basic GPSR and MORA.

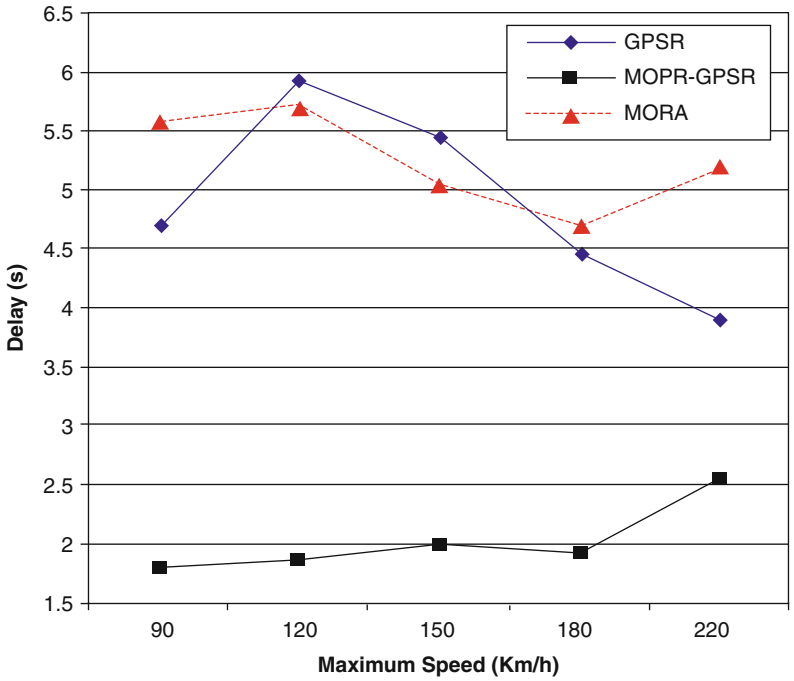
To evaluate the performance of any routing protocol in terms of routing overhead, it is important to



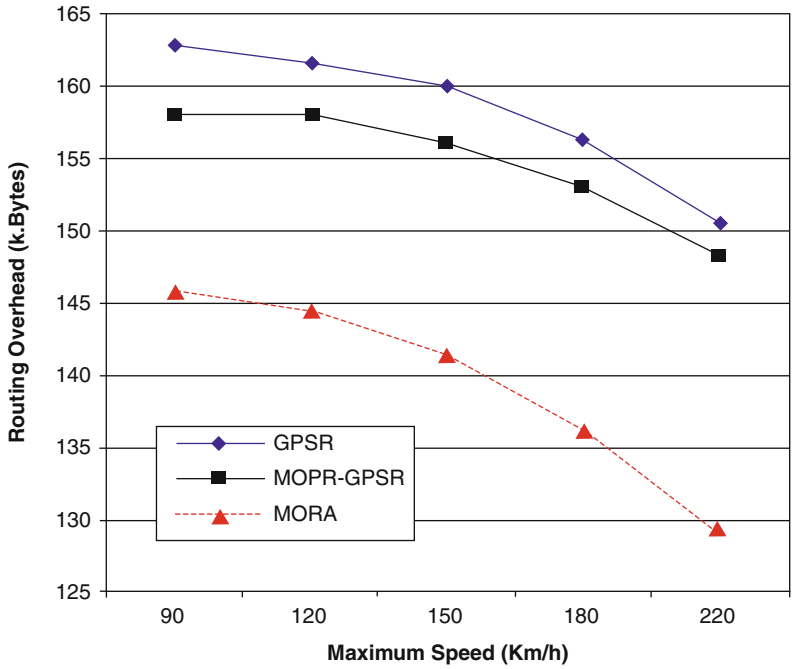
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 8
The highway scenario in simulations



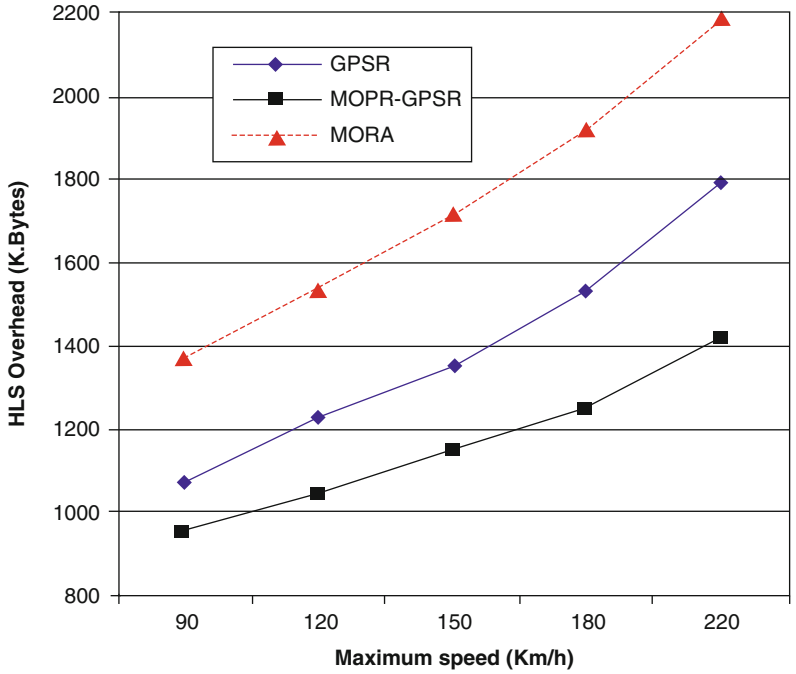
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 9
Packet delivery ratio comparison between GPSR, MOPR, and MORA



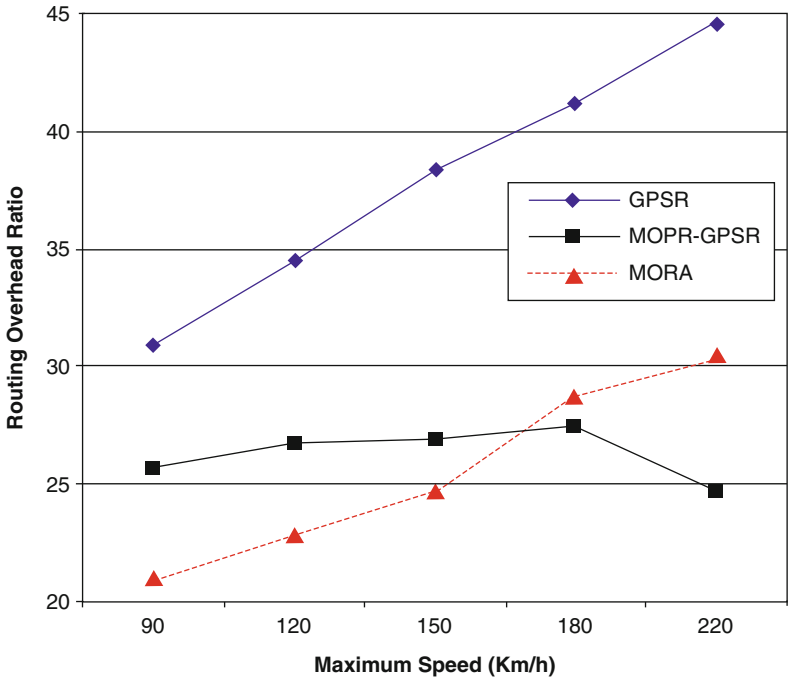
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 10
Delay comparison between GPSR, MOPR, and MORA



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 11
Routing overhead comparison between GPSR, MOPR, and MORA



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 12
HLS overhead comparison between GPSR, MOPR, and MORA



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 13
Routing overhead ratio comparison between GPSR, MOPR, and MORA

calculate the Routing Overhead Ratio (ROR). Figure 13 shows the ROR caused in the simulation network while taking into account only the routing overhead (i.e., without counting the HLS overhead). It is clearly shown that both MOPR and MORA improve the ROR when compared to basic GPSR. And MOPR shows an almost stable ROR compared to MORA which increases when the maximum speed increases.

The ROR is important, but in such kind of routing protocol, the global ROR overhead, while taking into account the HLS overhead as well, is more important. Figure 14 shows clearly how MOPR improves the network performance in terms of global ROR by about two times when compared to both basic GPSR and MORA.

All simulation results presented in this section show that MOPR improves the routing performances from all sides. Consequently, MOPR shows a big potential for position-based routing in VANETs.

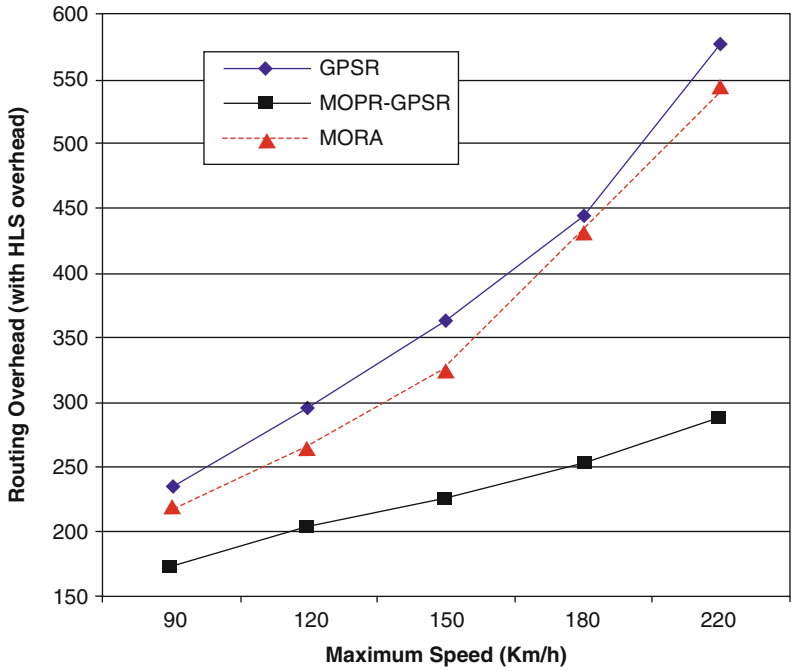
Strategies of Enhancement in City Scenario

Definition of the Subject A city scenario is much more complicated. Communication between cars driving in a city environment creates different challenges, largely due to the more complex geometry of the scenario. The challenges can be summarized as follows:

Geometric two-dimensionality: Cars change their movement direction all the time and can move at any relative angle to each other allowed by the street geometry. This weakens the correlation of the destination position to a suitable next hop.

Obstacles: A city is usually characterized by the presence of radio obstacles, which creates problems with position-based next hop selection because the nodes are not able to communicate whenever the line of sight between two nodes goes through an obstacle. And multipath propagation and complex obstacle surfaces create a much more complicated situation in reality.

Node density: the node density can be expected to be rather high with respect to the radio range, especially at "density hot spots" like junctions. Node density creates better ad hoc connectivity while it also poses a challenge to flooding mechanisms that need to be very efficient.



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 14
Global routing overhead (routing + HLS) ratio comparison between GPSR, MOPR, and MORA

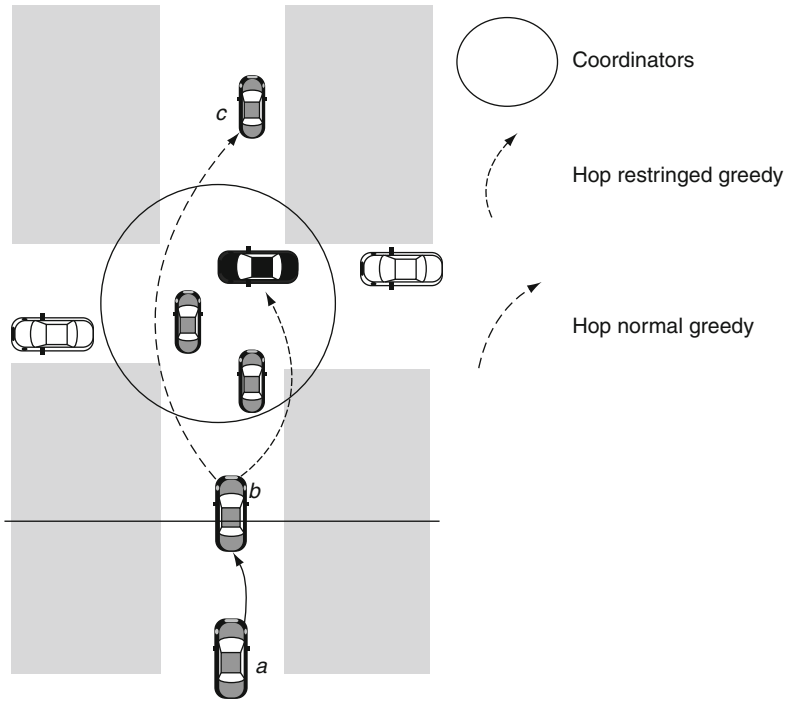
Low mobility: Compared to highway scenarios, nodes move at lower speeds, influenced by node density and are constrained by speed limits.

Greedy Perimeter Coordinator Routing

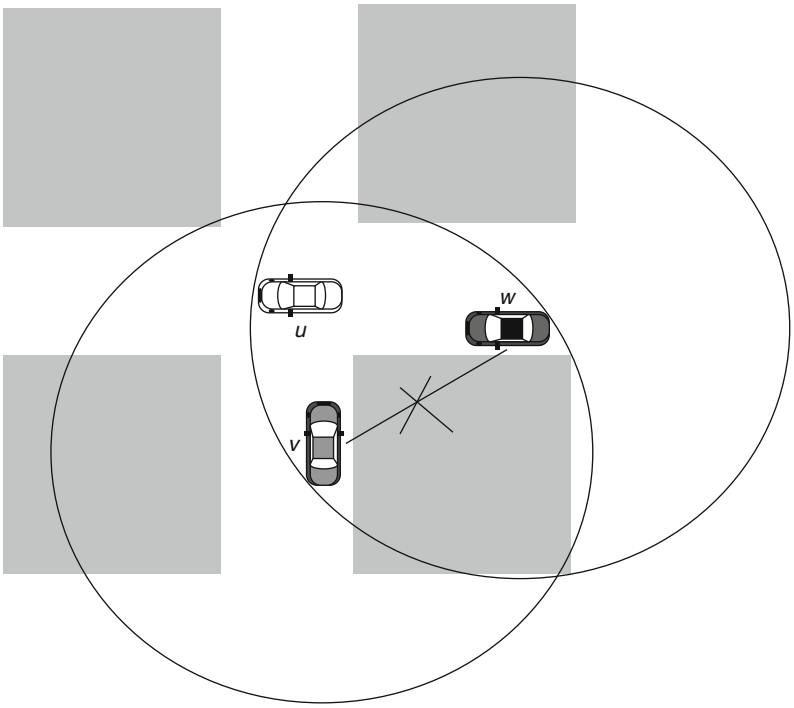
Greedy perimeter coordinator routing (GPCR) [7] does not use map information or densities in the streets, nor does it use the idea of a list of junctions that a packet must pass until it reaches the destination. It aims to avoid overhead. GPCR proposes a position-based algorithm. The main idea is to take into account the fact that the streets and junctions form a natural planar graph and, hence, it is possible to apply geographic routing directly. GPCR consists of two parts: restricted greedy routing and repair strategy.

Restricted greedy routing. Data packets should be routed along streets because they cannot get through buildings. The junctions are the only places where actual routing decisions are taken. Therefore, packets should always be forwarded to a node on a junction rather than being forwarded across a junction. This is illustrated in Fig. 15, where node *b* would forward the packet beyond the junction to node *c* if greedy forwarding is used. But by forwarding the packet to any of the nodes in the corner it finds an alternative path to the destination without getting stuck in a local optimum. A local optimum is produced when a forwarding node does not find a neighbor closer to the destination than itself. A node on a junction usually has more available options to route a message. Nodes that are located close to a junction are called coordinators. A coordinator broadcasts its role into its beacon packets. Thus, its neighbors will know its role when they have to forward these beacons. As a node must know whether it is a coordinator or not, two methods have been proposed to learn that.

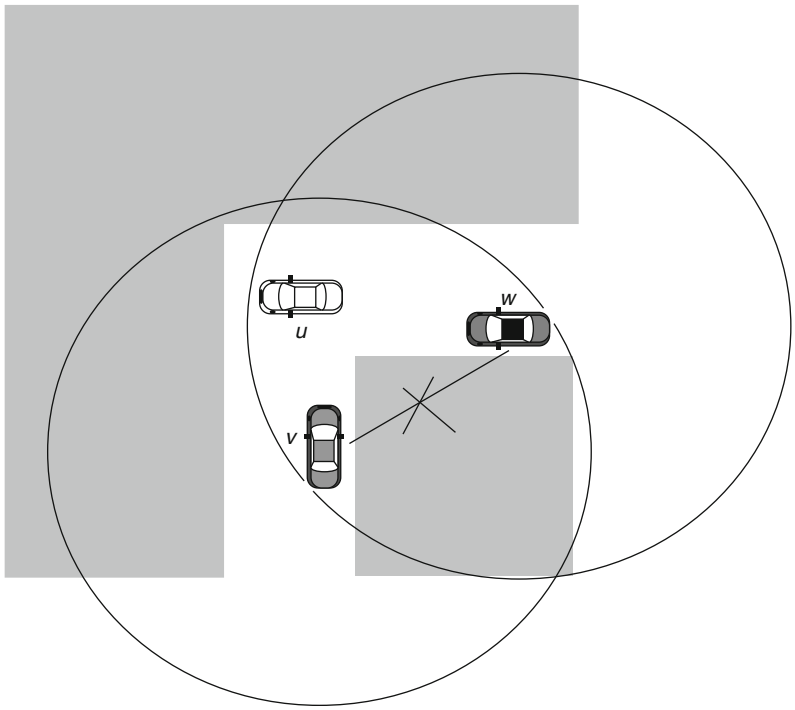
The first one consists of including neighbors' locations and identifiers in beacons, so that each node can have information about the two-hop distance. Then, a node is considered to be in a junction when it has two neighbors that are within transmission range to each other but do not list each other as neighbors. Figure 16 shows *v* and *w* are neighbors of node *u* but they do not list each other as neighbors. This method might have problems. In Fig. 17 the requirement is correct but



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 15
Preference coordinator nodes in GPCR



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 16
Discovery method of coordinators in GPCR



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 17
Failed discovery coordinator in GPCR

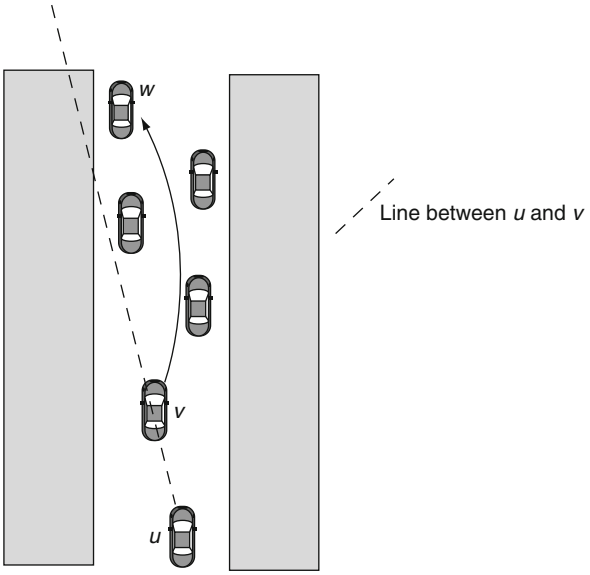
node u is not really a coordinator because it is located in a curve.

The second one consists of calculating the correlation coefficient (CC) with respect to the position of the neighbors. In this method, it is not necessary to include additional information into beacon messages. Let x_i and y_i be the (x, y) coordinates of a node i. Also let $\bar{x}$ and $\bar{y}$ be the mean of x-coordinates and y-coordinates, respectively. σ_{xy} indicates covariance of x and y. σ_x and σ_y indicates the standard deviation of x and y, respectively. Finally, the correlation coefficient is defined in Eq. 7:

$$\rho_{xy} = \left|\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right| = \left|\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}\right| \tag{7}$$

with $\rho_{xy} \in [0,1]$. If the value is close to 1 it indicates a linear coherence that is normally found when a vehicle is located in the middle of a street (Fig. 18). On the other hand, a value close to 0 might indicate a linear coherence and, hence, a vehicle can be determined to be located in a junction. By adjusting a threshold $\in$, a node can evaluate the correlation coefficient and assume with $\rho_{xy} \geq \in$ that it is located on a street and then the node is a coordinator. But if $\rho_{xy} \leq \in$, it can be concluded that the node is close to a junction. 0.9 is considered to be a good value for 2, but it may be arbitrary.

If the forwarding node is located on a street and not on a junction, the packet is forwarded along the street toward the next junction. To achieve this, the forwarding node draws a line between the forwarding node's predecessor and the forwarding node itself. The neighbors that approximate the extension of the line will be candidates. The farthest candidate node is selected. In Fig. 18, node w is selected based on the line between u and v. In the event that some candidate node is a coordinator it will be selected before any non-coordinator node. If there were more coordinators, one of them is randomly selected (Fig. 15). This prevents a packet from crossing a junction. Once a packet reaches a coordinator, a decision has to be made

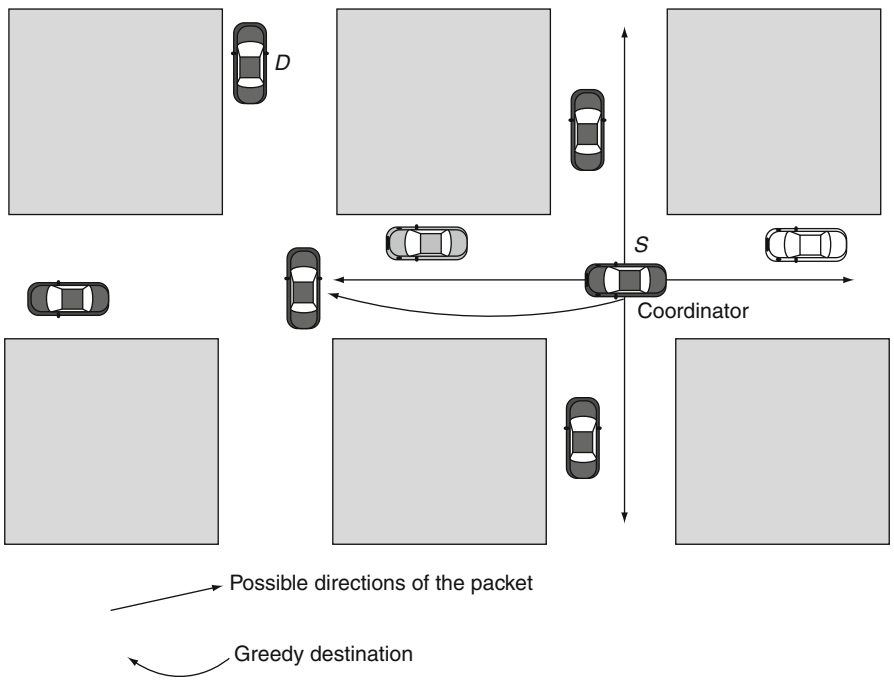


Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 18
Restricted greedy in a street

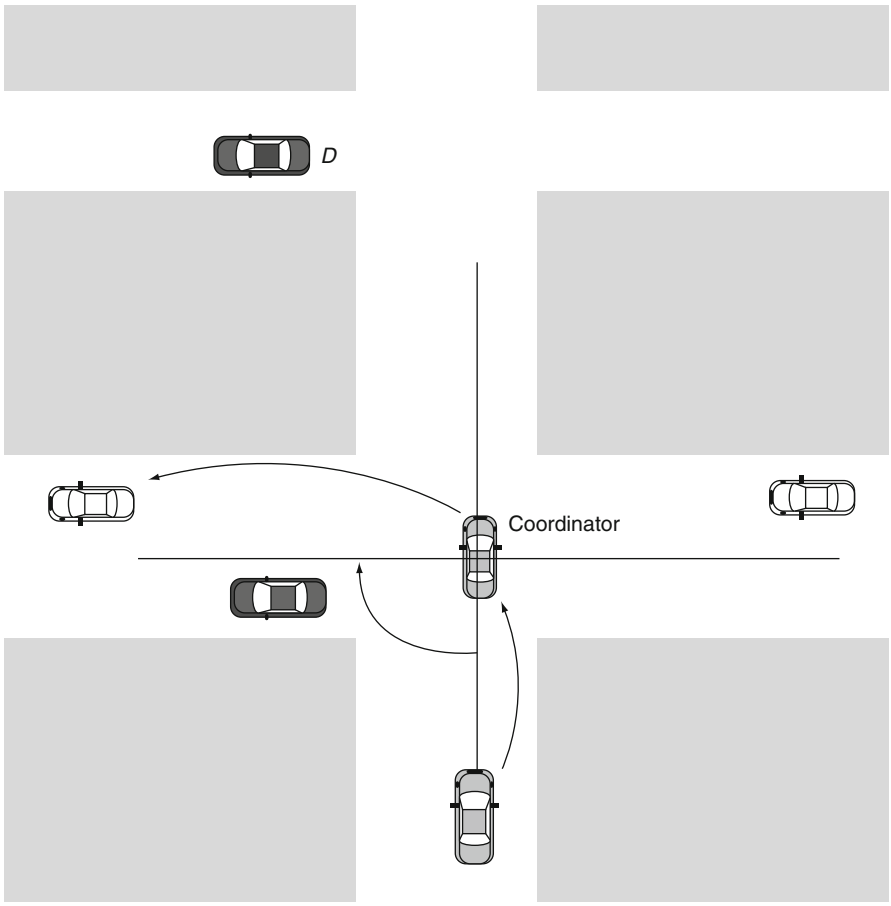
about the street that the packet should follow. This is done in a greedy mode: the neighbor with the largest progress toward the destination is chosen. This implies a decision on the street that the packet should follow (Fig. 19).

Repair strategy. Despite the new greedy routing model, there exists the risk that a packet gets stuck in a local optimum. Hence a repair strategy is required. The vehicle tries to infer the topology of the roads by applying the recovery strategy over the set of neighbors. If the forwarding node is a coordinator and the packet is in repair mode, then the node needs to determine which street the packet should follow next. To this end the right-hand rule [8] is applied (Fig. 20). Using the right-hand rule it chooses the street that is the next one counterclockwise from the street the packet has arrived on. But if the forwarding node with a packet in repair mode is not a coordinator, then the node applies restricted greedy routing.

In GPCR, there exists a risk that a packet could be forwarded back over the same street from which the packet has arrived. When a packet is being forwarded in repair mode and reaches a coordinator node, it applies



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 19
Restricted greedy in a street

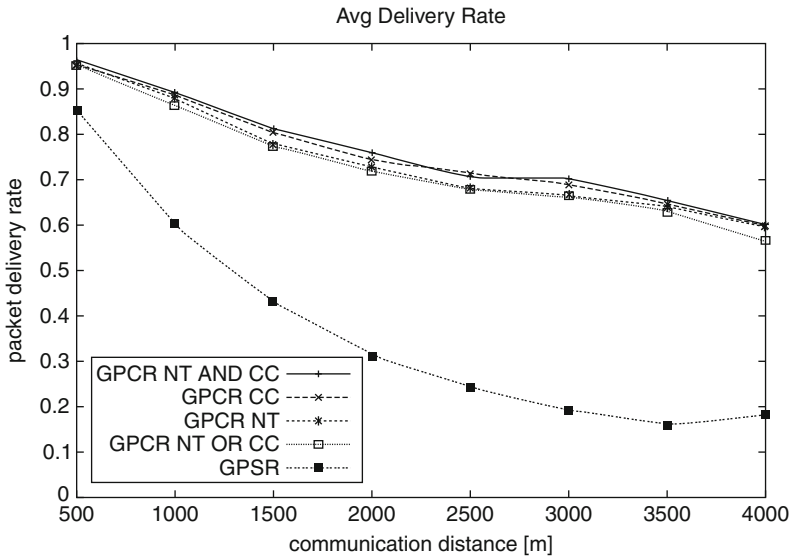


Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 20
Coordinator applying perimeter in GPCR

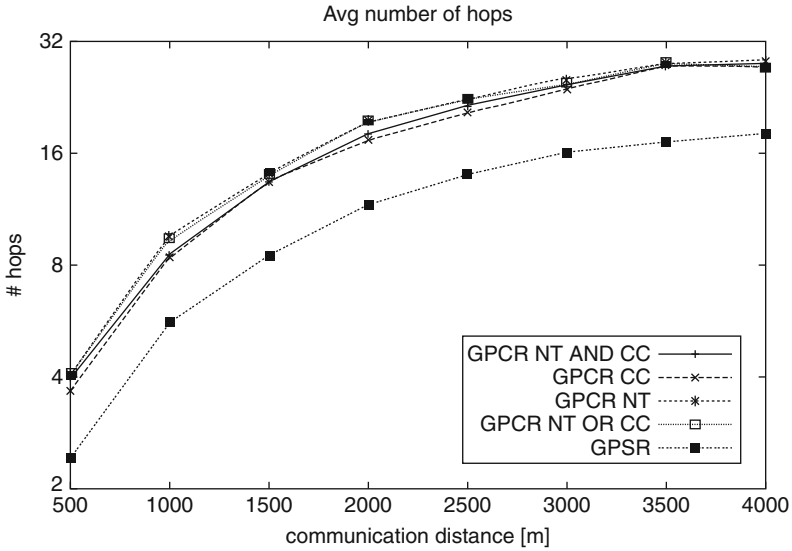
perimeter routing. In Fig. 21, it shows how if the node u applies the right-hand rule (in this case left-hand rule) from the line formed between nodes u and v, the coordinator chooses the node w as the next hop, instead of the node x which is located along the next street. Therefore, the packet does not turn the junction, but it remains on the same street.

Simulation of GPCR is implemented with the ns-2 simulator version ns-2.1b9a. In the simulation, a real city topology is used, which is a part of Berlin, Germany. The scenario consists of 955 cars (nodes) on 33 streets in an area of 6.25 × 3.45 km. The movement of the nodes was generated with a dedicated vehicular traffic simulator and represents a real-world movement pattern for this given scenario [9]. IEEE802.11 was used as MAC with a transmission rate of 2 Mbps. The transmission range was set to 500 m. Real-world tests with cars have shown this to be a reasonable value when using external antennas. For each simulation run ten sender–receiver pairs are randomly selected. Each pair exchanges 20 packets over 5 s. Figure 21 shows the achieved packet delivery rate versus the distance between the two communication partners and Fig. 22 shows the number of hops. The communication distance between two nodes is calculated as the minimal distance based on the street topology at the beginning of the communication.

Figure 21 also depicts how the delivery rate is influenced by the algorithms used for junction detection. It shows that calculating the correlation coefficient (CC) is slightly better than relying on the comparison of the neighbor tables (NT).



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 21
GPCR versus GPSR – delivery rate



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 22
GPCR versus GPSR – average number of hops

A compound decision consisting of the neighbor table comparison and correlation coefficient is also analyzed, concatenated by logical OR as well as by logical AND. The latter one outperforms the other approaches slightly but it does not come for free: the size of the beacon packets increases for each of the two approaches. Therefore, GPCR simply uses the correlation coefficient. In general, the study on achievable

packet delivery rate (Fig. 21) shows good results for the approach compared to GPSR. This improvement in performance comes at the expense of a higher average number of hops and a slight increase in latency. This increase in hop counts and latency is mainly caused by those packets that could not be delivered at all by GPSR and thus did not impact the hop-count and latency for GPSR.

Connectivity-Aware Routing

Connectivity-aware routing (CAR) [10] is a position-based routing scheme. The protocol is aimed at solving the problem of determining connected paths between source and destination nodes. VANETs' nodes present a high degree of mobility, and nodes cannot know the position of the rest of the vehicles due to several well-known scalability problems. This lack of information makes it impossible to determine a priori which streets have enough vehicles to allow messages to be routed through them.

CAR's algorithm is designed to deal with these problems, and to do that it is divided into three stages: (1) finding the location of the destination as well as a connected path to reach it from the source node, (2) using that path to relay messages, and (3) maintaining the connectivity of the path in spite of the changes in the topology due to the mobility of vehicles.

In the first stage, the source node broadcasts a route request message. The idea behind this initial broadcast is the following: The reception of, at least, one of these route request messages at the destination means that at least one connected path exists. The destination node answers the route request message with a response message including its current location so that the first problem is solved. But the source node also needs to know the path to reach the destination.

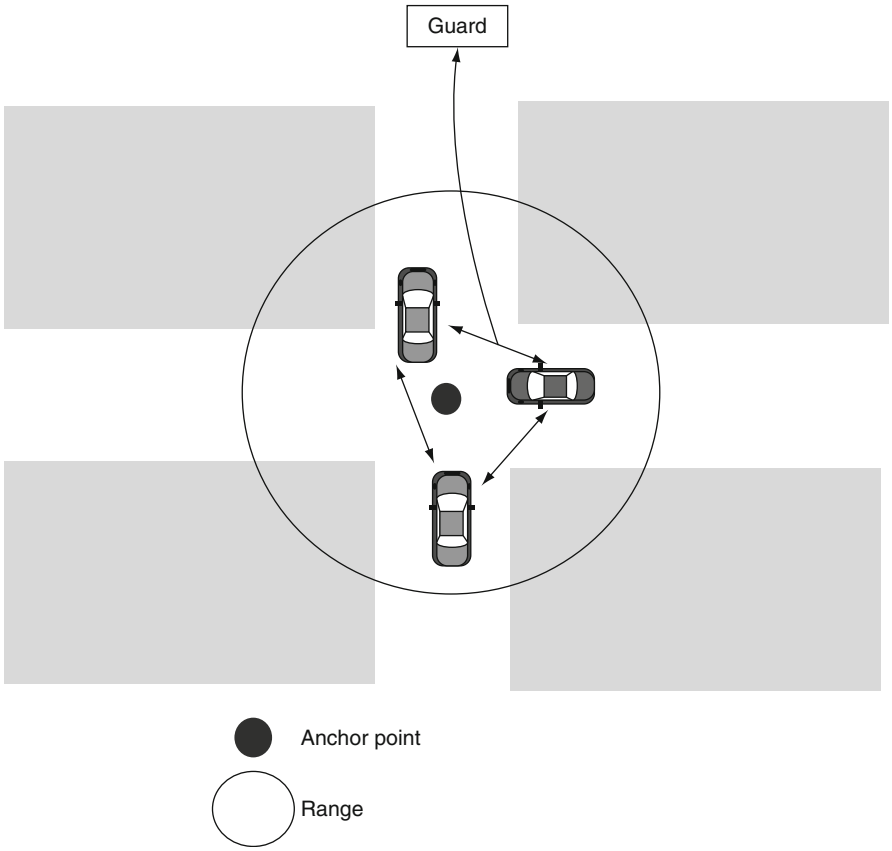
In CAR, it is proposed to include in the header of every route request message the list of junctions (called anchor points) traversed by that message in its way toward the destination. Thus, adding that list to the response message issued by the destination solves the second problem. Besides, nodes periodically transmit short messages including the issuer's identifier, location, and current velocity vector. These short messages (called beacons) keep the neighbor's tables updated. Moreover, the determination of junctions is made by means of comparing the direction of the vehicles. That is, a node determines if it is currently located at a junction when the angle between its velocity vector and one of the neighbors is not parallel. Two velocity vectors are parallel if the smallest angle between the vectors is less than α (equal to 18°). Otherwise the velocity vectors are nonparallel. Nodes that have neighbors with nonparallel velocity vectors identify themselves as being near a crossing or road curve and can serve as relays.

Additionally, to select not only a connected path between the source and the destination but also a short one, the destination does not respond immediately. Instead, it waits a predefined amount of time and then the path selected is the shortest one among those included in the different route request messages received. CAR uses the preferred group broadcast (PGB) [2] protocol to reduce as much as possible the overhead of flooding.

Once the source node has determined a path to reach the destination, data messages are routed geographically from an anchor node to the next one until the destination is reached. To do that, the source node uses a source routing approach. The full list of anchor nodes is included as a header in every data message transmitted. CAR uses the advance greedy forwarding (AGF) [2] algorithm to deliver messages between each pair of anchor nodes. In AGF, relay nodes select as next hop the neighbor located closest to the destination. In this case, it is the vehicle located closest to the next anchor point.

CAR defines the concept of "guards" (see Fig. 23) to help nodes determine if a message has reached a certain anchor point. A guard is a set of information tied to a geographical area. That area is defined by the location of the anchor point and a radius. Thus, guards contain both the location and the radius. Nodes create guards when they identify a new anchor point. A node creating a guard is the first one including it on its beacon messages. Nodes store the guards received during beaconing periods, but only nodes located inside the area defined in the guard retransmit it on their beacons.

A node receiving a message being routed toward the anchor point *a* can determine that the message has reached *a* by checking if it has already received a guard for *a*. Moreover, the mobility of vehicles causes



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 23
Three vehicles interchanging a guard about an anchor point.

constant topology changes. Thus, the connected path found at the beginning of a data transmission can become disconnected over time. To overcome this issue, CAR uses the guards to help maintain the connectivity of the path, or at least to dynamically auto-adjust it on the fly without resorting to a new route discovery process.

Concretely, it is assumed that there cannot be disconnection problems between anchor points, so that only the movement of the destination node represents an issue. Therefore, when a destination node changes its direction, then a new guard is generated including also the new velocity vector of the destination node. When a data message arrives to the old destination node's location, the guarding nodes (those interchanging that guard) can retransmit the packet toward the new estimated location of the destination. Of course, this assumption may not hold in general, which means that the protocol may fail to maintain the path connected.

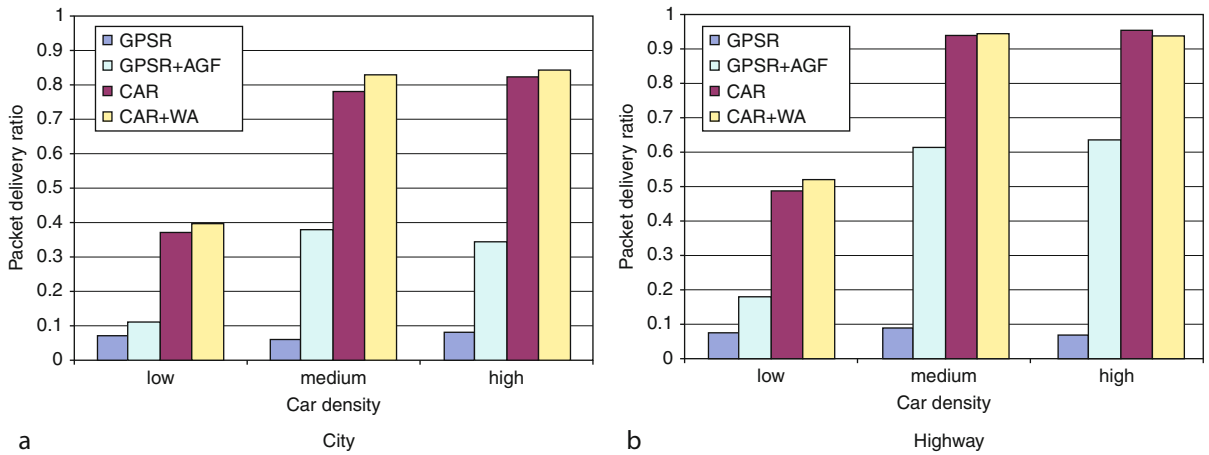
Finally, as CAR makes extensive use of beacons, an adaptive beaconing mechanism is proposed to reduce control overhead while keeping neighbor tables as accurate as possible, especially when the number of neighbors or their mobility makes them very unstable. The idea is to adapt the beaconing rate to the nodes density, so that the fewer the number of neighbors, the higher is the beaconing frequency. By this way, several drawbacks, such as wasted bandwidth, delaying of data packet, increased network congestion, coming from fixed period beacon strategy, can be mitigated.

In CAR, there are two possibilities for routing error to occur. First, the AGF algorithm may fail to forward a packet between two anchor points due to

(a) a temporary gap between vehicles (or raised interference level), such gaps may appear and disappear with time at any place on a road; or (b) long-term disconnections due to a suddenly closed road or an unusually big gap in the vehicular traffic. Second, a packet may reach the estimated destination position after passing the last anchor point but fails to find the destination there. The reason for that event could be that the destination changed direction but could not activate a guard due to lack of neighbors within communication range, or the guard was activated but later could not be retransmitted due to the same problem. CAR uses two approaches to handle these routing errors: (1) *Timeout algorithm with active waiting cycle* – One approach to tackle temporary gaps (or a raised interference level) is the use of timeout with packet buffering and an active waiting cycle. The forwarding node suspends the packet and periodically checks if the next hop neighbor has appeared. A long-term disconnection recovery algorithm should be invoked when a simple timeout approach failed. (2) *Walk-around error recovery* – If the AGF algorithm fails to find the destination at its estimated position (case 1), or the timeout algorithm could not find the next hop host (case 2), the node that detected the problem informs the source about the error and starts a local destination location discovery process. In case (1) the scope of this discovery is limited to half the number of anchor points in the old source–destination path. The broadcast is allowed to travel no more than one half of the old path length. In case (2) the scope is limited to the number of anchor points in the old path to the destination (from the current node) plus 50%. The same applies to the path length.

Simulations are made under version 2.28 of the ns-2 simulator with the probabilistic Shadowing model. The performances of CAR protocol without and with enabled walk-around error recovery (CAR + WA) are compared with GPSR and GPSR + AGF. Three different densities of nodes (low – less than 15 vehicles/km of road, medium – 30–40 vehicles/km, and high – more than 50 vehicles/km) are used in the following movement scenarios: highway (averaged over three different highway areas, ten sub-scenarios each for every density of vehicles) and city (averaged over three different city areas, ten sub-scenarios each for every density of vehicles) [2]. 20 CBR traffic sources with a sending rate of 4 packets/s are considered. Sources stop generating data packets 50 s before the simulation ends. Source/sink nodes stay inside the simulated area (do not leave the area and do not park) for the duration of the simulations (300 s).

Figure 24 shows packet delivery ratio for city and highway scenarios with different densities of vehicles. For all traffic densities, GPSR performs very poorly in the city scenario, with 5–7% of data packets delivered. Also, the advanced greedy forwarding algorithm (GPSR + AGF) shows moderate performance (up to



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 24
Packet delivery ratio

only 38% of data packets delivered), although performance is noticeably higher than for standard GPSR. Note that GPSR and GPSR + AGF use an idealized location service in the simulation: source nodes obtain the true location of destinations each time a data packet is originated. Despite the additional overhead to discover the real paths and to obtain destination coordinates, CAR and CAR + WA demonstrate much better results than GPSR + AGF. The highway scenarios are geographically less sophisticated than the city scenarios, thus all studied protocols show better PDR in highway areas. Again, CAR and CAR + WA outperform GPSR and GPSR + AGF, despite the need to obtain and maintain paths between source–destination pairs.

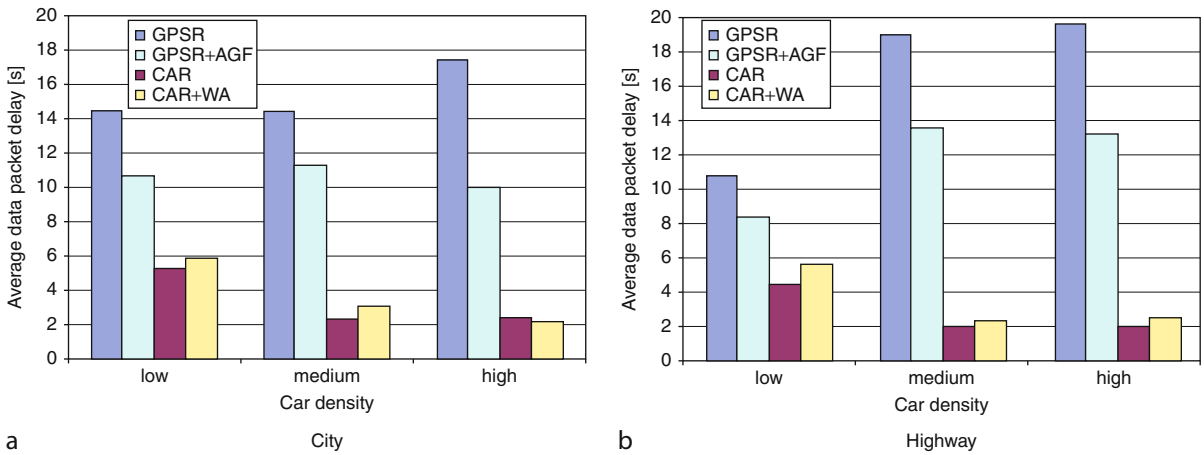
In terms of the average data packet delays (Fig. 25), the original GPSR and the GPSR + AGF are always worse than CAR and CAR + WA. For CAR, the route discovery process precedes every first data transmission to an unknown destination, this step adds to the delay of the first data packets. However, the average delay of the data packet for CAR and CAR + WA is much lower than for GPSR and GPSR + AGF. This result is a consequence of CAR's use of real connected paths between source and destination pairs, whereas GPSR and GPSR + AGF often fail due to local maximum resolution encountered by the perimeter mode. CAR easily tolerates short-term disconnections due to gaps or a temporary high interference level (e.g., frequent MAC collisions).

Figure 26 shows the total routing protocol overhead, measured in total number of routing packets sent network-wide during the entire simulation. For the CAR protocol, the overhead is presented as accumulative contribution of (1) beaconing, (2) path discoveries, and (3) path maintenance with the help of guards. The use of adaptive beaconing allows CAR to keep the average beaconing overhead from 1.5 to 3 times lower than the beaconing overhead of GPSR, without harming the performance.

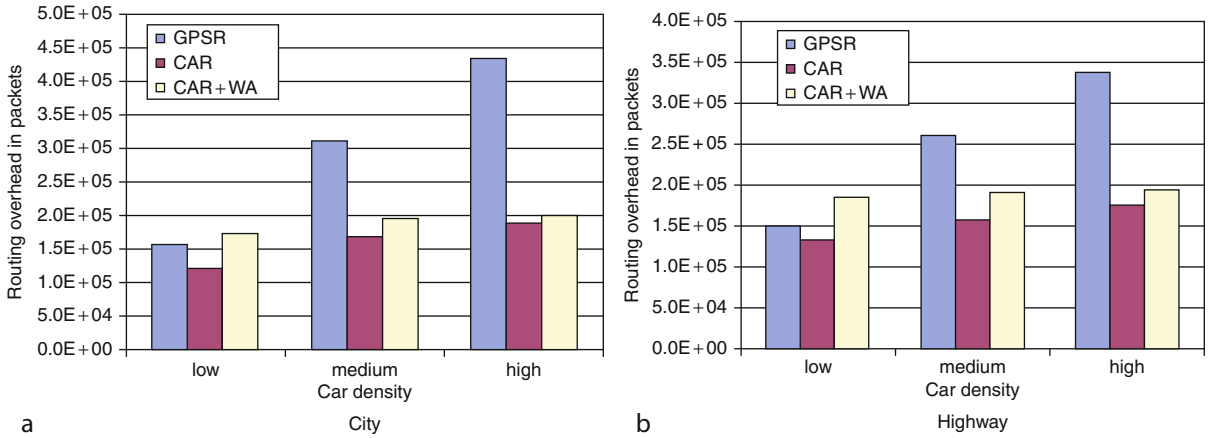
GpsrJ+

GpsrJ+ is a position-based routing protocol which consists of two modes, yet using a special form of greedy forwarding. As obstacles (e.g., buildings) block radio signals, packets may only be greedily forwarded along road segments as close to the destination as possible. Accordingly, the major directional decisions are made at junctions. When packets reach a local maximum, a point at which there is no node closer to the destination, the node switches to GpsrJ+'s recovery mode.

In recovery mode, packets are greedily backtracked along the perimeter of the roads. GpsrJ+ removes the unnecessary stop at a junction while keeping the efficient planarity of topological maps. It is not necessary to back forward in small steps through planarized links, for reasons that the general direction of the right-hand



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 25
Average delay of a data packet



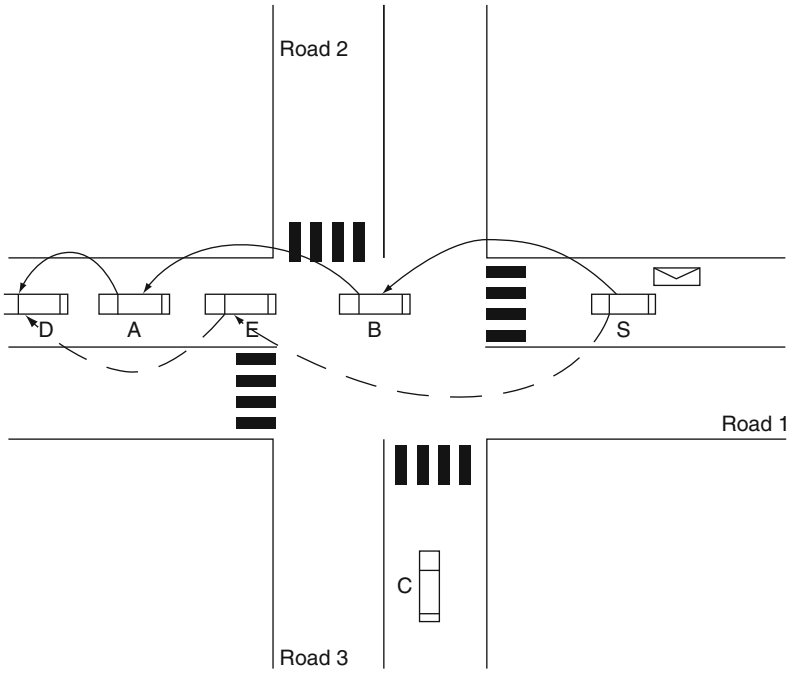
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 26
Routing overhead in packets

rule always results in the opposite direction of where packets were going before recovery, and the objective is to come back as fast as possible to a junction. Unlike GPCR, where packets must be sent to a junction node since junction nodes coordinate the next forwarding direction, GpsrJ+ lets nodes that have junction nodes as their neighbors predict on which road segment its junction nodes would forward packets onto, and thus may safely overpass them if not needed. GpsrJ+ uses the right-hand rule to determine the best direction (as opposed to final destination direction) and thereby the best forwarding node. That is, if the furthest node is in the same direction as the best direction, the best forwarding node is the furthest node; otherwise, the best forwarding node is a junction node. Figure 27 illustrates the advantage of prediction. The figure shows that GpsrJ+ can bypass the junction area and forward the packet to node *E* directly, yet GPCR forwards it to the junction node *B*, thus causing more transmissions.

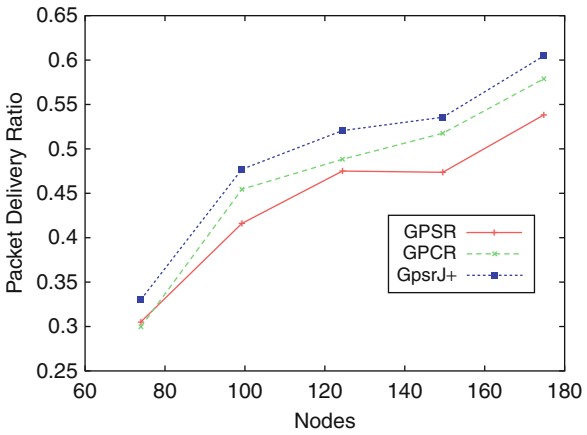
Moreover, GpsrJ+ uses a two-hop neighbor enhanced beaconing. In addition to the node's position in the beacon, each node also broadcasts the road segments that its neighbors are on. In the neighbor list, each node has its neighbor's location and the associated road segments on which its neighbor's neighbors are.

GpsrJ+ is analyzed by comparing it with GPSR and GPCR under Qualnet 2.95 simulator software, with 75 nodes up to 175 nodes, with a 25-node increment. Generally, GpsrJ+ improved recovery strategy brings significant results compared to GPSR and GPCR. In the simulation, an urban topology is employed, which is a user-defined Manhattan-grid of 1,500 × 1,500 m.

Figure 28 shows the packet delivery ratio (PDR) between GPSR, GPCR, and GpsrJ+. Clearly, taking aggressive hops in the recovery mode along the perimeter improves the PDR. This is further verified by the fewer hops GpsrJ+ needs compared to GPSR as shown in Fig. 29a. A higher number of hops imply an increased probability of channel contention; therefore, there is a higher probability that a packet gets dropped along the way. Although GPCR and GpsrJ+ stop at each junction node in greedy mode, this is not sufficient to increase the hop count dramatically. The total hop count of GPCR and GpsrJ+ is still lower than that of GPSR. Figure 29b shows the number of hops a packet experiences before being dropped. GPSR's failed hop is twice as much as GPCR and GpsrJ+. This is consistent with that planarization of nodes produces too many hops. The undeliverable packets, as a result of disconnections between the source and destination, engage in perimeter forwarding most of the time and explore all possible perimeters in a limited way caused by planarization. Since more nodes are involved in forwarding, there is a lot of resource wastage. The situation worsens for undeliverable packets as they create a loop and the same route formed by the same



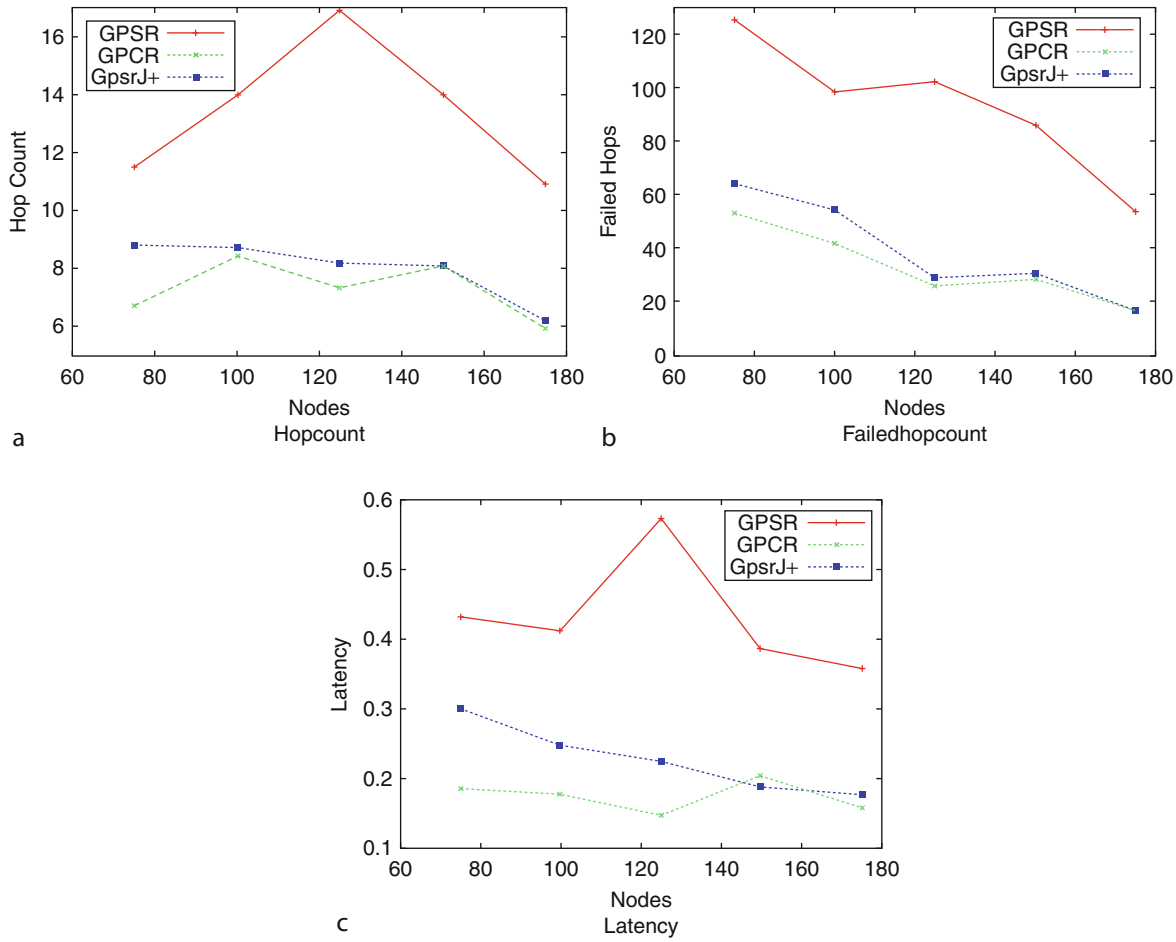
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 27
Dashed arrows are GpsrJ+ and *solid arrows* are GPCR



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 28
PDR among GPSR, GPCR, and GpsrJ+

nodes in the same hops is visited again. In summary, the inefficiency of node planarization strategies in urban vehicular scenarios to forward packets in perimeter mode not only affects the delivery ratio but also impacts the hop count and network resources as packets stay longer in the network before being dropped.

Figure 28 also shows that GpsrJ+ possesses a higher PDR than GPCR thanks to prediction. The slight increase in hop count and latency in Fig. 29a and c, respectively, is the result of packets that do not get delivered to the destination and thus do not contribute to GPCR's hop count and latency. The reason is that the smoother decrease in hop count in GpsrJ+ compared to GPCR is due to the fact that nodes do not necessarily have to go through junction nodes, which might be heavily used for forwarding in GPCR. Consequently, the interference and collision of multiple packet transmission cause packets either to be dropped or to be forwarded on a longer route. The slight increase in failed hops in GpsrJ+ compared to GPCR in Fig. 29b illustrates a longer expectancy of packets as GpsrJ+ makes a better effort to deliver them. Once again, the ability not to rely on junction nodes that get flooded with traffic prolongs the life expectancy of a packet before it gets dropped. The improved PDR in GpsrJ+ also brings in the advantage of the fraction of times a packet travels in greedy mode.



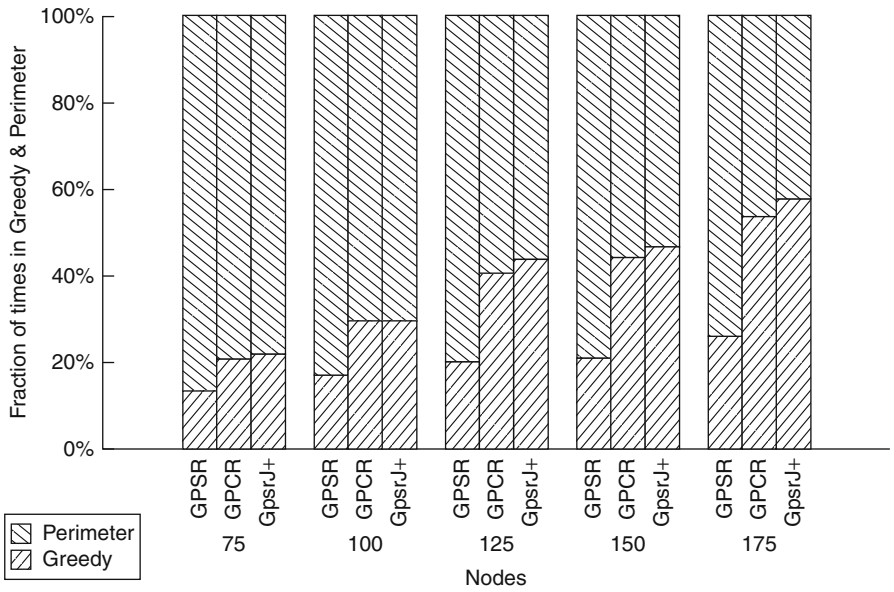
Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 29
Hop count and latency among GPSR, GPCR, and GpsrJ+

Figure 30 indicates that GpsrJ+ is in greedy mode a higher fraction of time than GPCR, and implies that GpsrJ+ minimizes the number of times a packet gets into a local maximum and maximizes the number of times a packet gets out of a local maximum.

Vehicle-Assisted Data Delivery

Zhao and Cao [11] proposed several vehicle-assisted data delivery (VADD) protocols. All of them share the idea of storing and forwarding data packets. That is, nodes can decide to keep the message until a more promising neighbor appears on their coverage range, but trying always to forward them as soon as possible. Additionally, decisions about which streets must be followed by the packet are made using vehicle and road information such as current speed, distance to the next junction, and maximum speed allowed. These routing decisions are dynamically taken at junctions because the authors state that precomputed optimal paths used by other protocols might rapidly lose their optimality due to the unpredictable nature of VANETs.

In VADD, the main goal is to select the path with the smallest packet delivery delay. The behavior of the protocol depends on the location of the node holding the message. Two cases are considered: when nodes routing the message are located in the middle of a road and when they are located in a junction. The first case (also called routing in straight way) presents less alternatives: forwarding the packet toward the next junction or to



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 30
Fraction of times in greedy and perimeter

the previous one. However, the second case (also called routing in intersections) is much more complicated because at junctions, the routing decision must consider the different roads, so that the number of options is higher.

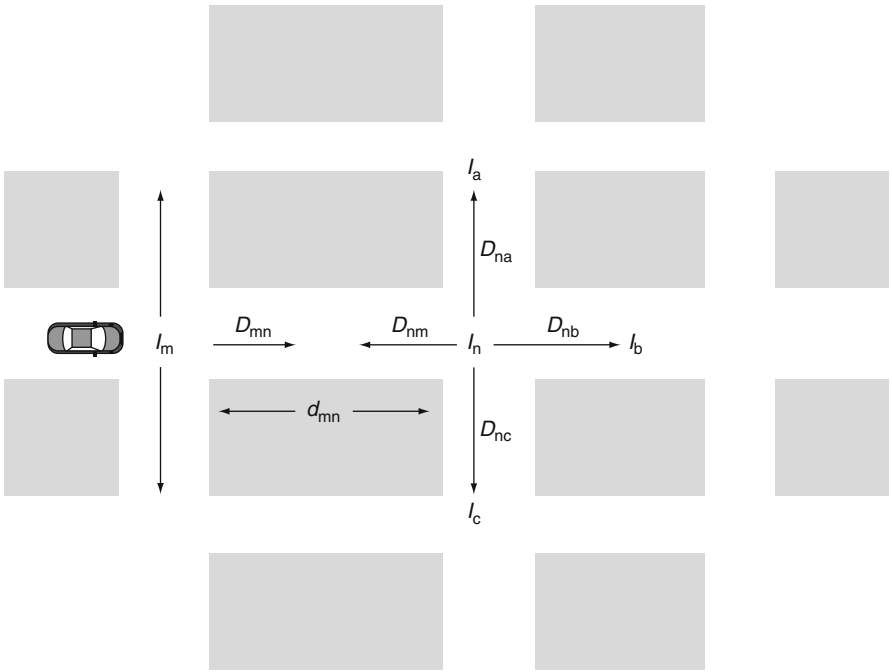
Both cases use the same approach, determining the next road the message must follow, and then selecting the next relay among the current neighbors. In VADD, a common way of determining the next road is proposed, while the determination of the next hop remains different. Concretely, the outgoing road with the lowest estimated delay would be selected. In the "straightway" case, there are two possible outgoing roads, the two segments of the roads in which the node divides the current road. In the "intersection" case, each road starting in that junction represents a different option.

The authors call roads to the street segments delimited by two consecutive junctions. The estimation of the delay of routing a message through a certain road takes into account the road's length, its maximum speed, the mean traffic density, and other traffic-specific parameters. But, to estimate the delay to the destination, the authors also incorporate the estimated delay of the next possible roads along with the probability of choosing them. Figure 31 depicts the delay model used in VADD. A car located near intersection I_m computes the delay for the road between I_m and I_n (D_{mn}) accounting also the estimated delay of choosing the road between I_n and I_a, the one between I_n and I_b, or the one between I_n and I_c.

To estimate message delays for the different roads, the authors propose to solve an $n \times n$ linear equation system using the Gaussian elimination algorithm [$\Theta(n^3)$], n being the number of junctions. To limit the complexity of this computation, a boundary area around the current location is defined, so that only the junctions inside that area are considered in the equation system.

Once the next road has been selected, it is time to determine which neighbor must be the next relay. In the "straightway" the decision is simple, the one located closest to the next junction according to the next road selected. In this case, the next junction can be the next one in the direction of the current vehicle or the one the vehicle has just passed by. In both cases, the packet is stored only if no neighbors are available at the moment.

The "intersection" case is more complex. Obviously, if no neighbor is available, or every outgoing road has a longer estimated delay than the current one, the decision taken consists of storing the message waiting for the next forwarding opportunity.



Vehicular Ad Hoc Networks, Enhanced GPSR and Beacon-Assist Geographic Forwarding in. Figure 31
Example of VADD model

Additionally, the authors propose three alternative ways to select the next forwarder when more than one candidate neighbor is available:

Location first: The node located closest to the next selected junction in the most promising road is chosen. This scheme presents routing loops.

Direction first: The node located closest to the next selected junction in the most promising road among the ones moving in the right direction.

Hybrid: Location first is applied unless a cycle is detected, in that case direction first scheme is used. This scheme seems a little bit unrealistic due to the difficulty of detecting routing loops.

VADD's main drawbacks are its complexity and difficulty of parameterization. The size of the bounding area is by far the most important parameter and, at the same time, it is responsible for the complexity of the computation needed at every node. Determining a value for this parameter to achieve a good trade-off between computational complexity and accuracy can be a hard task. Additionally, the authors claim that their hybrid scheme achieves the best performance; however, it is not clear how to implement this scheme due to the difficulty of detecting cycles.

Future Directions

The ways of strategies to enhance or extend performance of GPSR routing protocol have been introduced. The density of vehicles in each possible route as well as the travel direction and movement of the vehicles have been taken into account to improve the forwarding algorithm performance, and thus increasing the packets delivery ratio. Beaconing strategy has also been modified by these facts, and finally decreasing the route overhead.

Although data dissemination and routing have been extensively addressed, many unique characteristics of VANET together with the diversity in promising applications offer newer research challenges. Authors in [12, 13] have focused on the problems of radio obstacle during data transmission in city scenario.

They use an intersection-based approach to forward packets along successions of road from the source to destination in order to find robust and optimal routes. And the digital map information is used in these routing strategies to get information of intersections.

A VANET may exhibit a bipolar behavior, i.e., the network can either be fully connected or sparsely connected depending on the time of day or on the market penetration rate of the wireless communication devices. Authors in [14] have mentioned the issue when routing is within sparse vehicular ad hoc wireless networks. Simulation results show that the network re-healing time can vary from a few seconds to several minutes. This suggests a new ad hoc routing protocol will be needed as current routing protocol may not work with such long re-healing times for most of them are designed under an assumption that the networks are full and well connected.

Bibliography

Primary Literature

1. Naumov V, Baumann R, Gross T (2006) An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces. In: Proceedings of ACM MOBIHOC'06, Florence, pp 108–119
2. Lochert C et al. (2003) A routing strategy for vehicular ad hoc networks in city environments. In: IVS'03, pp 156–161
3. Menouar H, Lenardi M, Filali F (2005) A movement prediction-based routing protocol for vehicle-to-vehicle communications. In: First international vehicle-to-vehicle communications workshop, co-located with MobiQuitous 2005, V2VCOM 2005, San Diego
4. Granelli F, Boato G, Kliazovich D (2006) MORA: a movement-based routing algorithm for vehicle ad hoc networks. In: IEEE workshop on automotive networking and applications (AutoNet 2006), San Francisco
5. Menouar H, Lenardi M, Filali F (2006) An intelligent movement-based routing for VANETs. In: ITS world congress 2006, London
6. Kieß W, Füßler H, Widmer J, Mauve M (2004) Hierarchical location service for mobile ad-hoc networks. Proc ACM SIGMOBILE Mobile Comput Commun Rev 8:47–58
7. Christian L, Martin M (2005) Geographic routing in city scenarios. ACM SIGMOBILE Mobile Comput Commun Rev Arch 9(1):69–72
8. Frey H, Stojmenovic I (2006) On delivery guarantees of face and combined greedy face routing in ad hoc and sensor networks. In: Proceedings of the ACM annual international conference on mobile computing and networking (Mobicom), Los Angeles
9. Lochert C, Hartenstein H, Tian J, Füßler H, Hermann D, Mauve M (2003) A routing strategy for vehicular ad-hoc networks in city environments. In: Proceedings of IEEE IV'03, Columbus, pp 156–161
10. Naumov V, Gross TR (2007) Connectivity-aware routing (CAR) in vehicular ad-hoc networks. In: Proceedings of 26th IEEE international conference on computer communications (INFOCOM'07), Anchorage, pp 1919–1927
11. Zhao J, Cao G (2006) VADD: vehicle-assisted data delivery in vehicular ad hoc networks. In: Proceedings of 25th IEEE international conference on computer communications (INFOCOM'06), Barcelona
12. Nzouonta J, Rajgure N, Wang G, Borcea C (2009) VANET routing on city roads using real-time vehicular traffic information. IEEE Trans Veh Technol 58(7):3609–3626
13. Jerbi M, Senouci S-M, Rasheed T, Ghamri-Doudane Y (2009) Towards efficient geographic routing in urban vehicular networks. IEEE Trans Veh Technol 58(9):1–1
14. Wisitpongphan N, Bai F, Mudalige P, Sadekar V, Tonguz O (2007) Routing in sparse vehicular ad hoc wireless networks. IEEE J Sel Area Commun 25(8):1538–1556

Books and Reviews

Lee KC, Haerri J, Uichin L, Gerla M (2007) Enhanced perimeter routing for geographic forwarding protocols in urban vehicular scenarios. In: 2007 IEEE Globecom workshops, 26–30 Nov 2007, pp 1–10

Vitrification of Waste and Reuse of Waste-Derived Glass

Enrico Bernardo, Giovanni Scarinci, Paolo Colombo
Department of Mechanical Engineering – Materials Division, University of Padova, Padova, Italy

Article Outline

Glossary

Crystallization Crystallization is the formation of crystals (i.e., a solid phase possessing short-, middle-, and long-range order) from a liquid or a solid. It occurs via a process consisting of two steps: nucleation and crystal growth. During nucleation, the structural units (atoms, ions, or molecules) spontaneously arrange themselves according to a regular geometry, which is specific for the crystal phase being formed. If this cluster, typically of the order of a few nanometers, has reached a critical size, which depends on the operating conditions (temperature, supersaturation, etc.), it becomes thermodynamically stable. The crystal growth is the subsequent growth of the nuclei that succeed in achieving the critical cluster size. In a crystal, the constituents are arranged in a defined and periodic manner (unit cell) that defines the crystal structure.

Durability Durability is the ability of a substance (or a structure) to withstand the interaction with the surrounding environment without detrimental effects for its properties. Highly durable materials can be subjected to a variety of chemical attacks without significant release of constituents (leaching) or decrease in their physical-mechanical characteristics.

Glass A glass is an inorganic solid material whose structure does not possess any middle- or long-range order. This condition is referred to as an amorphous (in comparison to crystalline) phase, featuring the distinctive phenomenon of glass transition. The most common glasses are based on oxides, and their composition may vary in a very wide range of values. A formal definition of glass is that it is an inorganic product of fusion, which has been cooled through its glass transition to the solid state without crystallizing.

Glass-ceramic A glass-ceramic contains both an amorphous phase and one or more crystalline phases. They are produced by a so-called, deliberate "controlled crystallization" process, which involves keeping the material for a certain time at temperatures at which the crystals first nucleate and then grow within the amorphous (glassy) matrix. In comparison to glasses, glass-ceramics possess a wider ranging set of thermo-mechanical properties.

Hazardous waste Waste are unwanted or unusable materials, which derive from natural or human (civil and industrial) processes. Hazardous waste are waste that pose substantial or potential threats to public health or the environment, and they can be flammable, radioactive, corrosive, toxic or have a genetic, carcinogenic, mutagenic, and teratogenic potential.

Vitrification Vitrification is the transformation of a substance into a glass, typically accomplished via a process involving the formation a liquid phase at high temperature, in the presence of an adequate content of vitrifying oxides (i.e., the so-called "network forming oxides," normally SiO_2, B_2O_3, P_2O_5). This subsequently cools to a solid without the formation of any crystalline phases. In this context, the substance of interest is hazardous waste, with the addition or not of additional glass-forming raw materials.

Definition

In the context of various technologies for the disposal of waste material, vitrification has proved to be the safest technology for the treatment and remediation of noncombustible hazardous waste. Vitrification is a process comprising the thermal destruction of waste. It is performed by melting the waste at high temperature, if necessary with the addition of glass-forming oxides (notably silica), and then cooling the melt quickly enough to prevent crystallization. In this process, the organic fraction is decomposed and burned off, whereas the inorganic fraction is stabilized by embedding its constituents at the atomic level within the glass network. The ability of glass to incorporate nearly all the elements of the periodic table into its structure is of fundamental importance: if formulated with an adequate composition, the resulting glass features a high chemical inertness, so that it can be landfilled without any particular concerns but it can also be recycled as secondary raw material for the manufacturing of new products.

In the early 1960s, it was first recognized that the immobilization of high-level radioactive waste (HLW) in glass was the most appropriate treatment to obtain an extremely durable and long-term stable waste form. After the establishment of the first vitrification plant of

Marcoule (France), more advanced vitrification systems have been developed and used in the USA, the UK, France, Germany, Belgium, Japan, and Russia. At present, vitrification is internationally still accepted as the best demonstrated available technology for the treatment of HLW [1, 2]. Given the excellent results achieved by HLW vitrification, in the 1980s and 1990s the technology of vitrification was gradually extended to other types of hazardous waste, and various types of melters – such as Joule-heated melters, electric arc furnaces, and plasma torch melters – were tested and commercialized. In Japan, because of the high cost and scarcity of available landfill sites, vitrification of municipal solid waste (MSW) incinerator residues (bottom ash and fly ash, containing hazardous substances such as heavy metals and dioxins) was developed in the 1980s, and has been in operation in a great number of MSW incineration facilities [3]. By this process, dioxins and furans in the residues are decomposed in the furnace at a temperature of approximately 1,400°C, and metallic compounds are stabilized in the structure of the product. In the USA, vitrification technologies were successfully commercialized [4–6] in the same period for the inertization of high-volume waste, including municipal sludge, paper sludge, soils, and dredged sediments. In this way, the need for disposal was eliminated and a glass aggregate was produced, marketable for many construction uses. During the 1980s, in situ vitrification earth-melting technology was developed by Battelle Memorial Institute [7] for DOE (US Department of Energy) and commercialized in the USA as GeoMelt^{TM} technology [8]. This method can be utilized for treatment of soils contaminated with high concentrations of both organic and inorganic polluting substances, and is the preferred methodology when there is need to avoid the risk connected with the excavation of the waste. Many applications of the in situ vitrification technology were carried out in the USA and Australia for the decontamination of polluted sites and soils. In the last years, plants based on the plasma torch melting technology have been realized in France for asbestos and fly ash vitrification [9].

Despite the soundness of vitrification technology, confirmed by numerous scientific studies and experimental tests, it has found difficulties in establishing itself. In fact, a cost analysis for vitrification [10], in comparison to other inertization options such as chemical processes or solidification/stabilization treatments, has shown that limited opportunity exists for a viable application of this process. An extensive use of vitrification can be carried out only when the absolute necessity of environmental safety has priority over cost, as is the case for radioactive waste. However, if the glassy material obtained by waste vitrification instead of being landfilled or used as inert aggregate for road foundations or other low value applications will be transformed into value added and competitive products, it will be possible not only to offset its production costs but also to make a profit. In the past years, a lot of experimental research has demonstrated the feasibility of such a transformation, with the realization of glass-ceramics, cellular glass, glass fibers, and other products from a variety of inorganic hazardous waste. Furthermore, recent studies have highlighted the possibility of utilizing the glass obtained from waste as a secondary raw material in manufacturing processes of ceramic articles of large consumption, such as bricks or paving tiles.

Introduction

Waste can be, in a first approximation, divided in two main categories: organic and inorganic waste, and each category includes both hazardous and nonhazardous waste. Organic waste are combustible, and today incineration of organic waste of all kinds (hazardous and not-hazardous) is the most common destruction treatment throughout the world, as it is easy to set up and operate and the heat generated can be utilized. A large variety of hazardous organic waste, including pesticides, polychlorinated biphenyls (PCBs), and persistent organic pollutants (POPs), is at present destroyed with very high efficiency by dedicated incinerators [11]. Inorganic hazardous waste derive mostly from many types of industrial processes, particularly in the metallurgical, steel, and chemicals production areas, but can come also from the demolition of buildings and civil infrastructures (realized, e.g., with cementitious materials containing asbestos), or form as a residue in many combustion processes, as the bottom and fly ashes produced by municipal solid waste (MSW) incineration plants. They require to be rendered inert before their landfill disposal or recovery, and their inertization can be achieved through chemical and physical processes such as: (1) stabilization using chemical agents, (2) extraction using acid or other solvents, (3) calcination at high temperature, (4) stabilization-solidification,

and (5) vitrification. The last two processes immobilize pollutants in a suitable matrix that indefinitely prevents their release into the environment upon disposal or reutilization. The goal is therefore that of reducing the polluting potential and the hazardousness of the waste, making them suitable for subsequent treatments (landfill disposal or recovery).

Stabilization-solidification processes include two stages: the first (stabilization) decreases the waste hazardousness by converting the chemical contaminants in a less soluble, less mobile and less toxic form; the second (solidification) only affects the physical state of the waste, turning them into a solid material with high structural and chemical integrity, thereby decreasing the mobility of pollutants. These processes can be carried out using thermoplastic or thermosetting polymers (which are however not frequently used because of their cost), or inorganic reagents such as cement, lime, or clay. However, in the latter case they increase the volume of waste, do not thermally decompose hazardous organics and do not give sufficient long-term guarantees in terms of the release of pollutants in the leachate [12].

Vitrification consists of the immobilization of hazardous waste, predominantly inorganic or with a low content of organics, in an inert amorphous matrix in the form of oxides, which, soluble in the molten glass at high temperatures, are then homogeneously incorporated into the vitreous structure following the cooling of the melt. It is therefore possible to successfully exploit one of the main characteristics of glass, which is to be a material characterized by high chemical stability that can contain a huge variety of pollutants in the form of oxides [12]. Hazardous waste can be vitrified by melting them and then cooling the melt at a rate sufficient to avoid crystallization. The waste, however, do not necessarily have to be intrinsically vitrifiable: if its composition does not contain enough glass-forming oxides (notably silica), it can be appropriately corrected through the use – to keep cost down – of vitrifying substances which are themselves waste, such as glass cullet or residues from the processing of feldspars. Glass cullet is particularly useful also as a low-cost flux to reduce the glass working temperature when this is too high, thereby helping to limit the cost of the process. Given the large quantity of thermal or electrical energy needed to melt the waste, the vitrification process is more expensive than others. Nevertheless, it is becoming increasingly important considering the rising cost of landfill, the exhaustion of available landfill sites and the hostility of the people toward the opening of new ones. Furthermore, the tightening of controls and sanctions against illegal disposal procedures and the enactment of increasingly stringent legal standards – which require preventive inertization of hazardous wastes – have given a significant boost to the technology of vitrification in all the most advanced countries.

Advantages and Prospects of Vitrification

The main advantages of vitrification can be summarized as follows:

1. Flexibility of the process, which allows to treat many types of waste, such as sludge, contaminated soil, ash, slag from hazardous processing, wet and dry solids (including asbestos, whose microfibers are destroyed only by heat), in large and variable proportions, often without the need for preliminary treatment
2. Destruction of all organics (including the most toxic substances such as dioxins and furans) with an efficiency exceeding 99.99%
3. Complete immobilization of the hazardous inorganic substances (such as heavy metals, radioactive elements, etc.) within the glassy network in ionic form
4. Substantial reduction in volume of the treated waste (from 20% to 97%, depending on the type of waste)
5. Good mechanical and thermal properties of the vitreous product
6. Excellent chemical stability and durability of the product, that is, resistance against attack from water or other chemical agents. Consequently, low environmental impact and possibility of landfill disposal without any problem, because any inorganic contaminant is retained permanently (any leakage of contaminants is so slow that no detectable adverse environmental effects are produced)
7. Well-established technology

The certainty that vitrification is the best available technology to ensure that the inertized waste do not

release toxic substances stems from the fact that it is the only process that has been adopted, since more than 40 years, for the long-term inertization of high-level radioactive waste in all countries that produce nuclear energy. Nevertheless, vitrification has been slow to become established in the field of hazardous waste inertization because it is expensive (for the high consumption of energy required to produce the oxide melt). Therefore, cheaper and easier solutions have been preferred, such as the immobilization in cementitious matrices (albeit less safe in the long term), the disposal in special landfills, or even illegal disposal methods (used for unspecified, but certainly very significant quantities of waste). However, if the glass obtained by vitrification, instead of simply being disposed of as an inert material or used for low value applications (e.g., road foundations, reinforcing filler for asphalt, drainage aggregates, the containment or consolidation of buildings and constructions, etc.), is transformed into value-added and competitive products, its intrinsic processing cost will be remunerated and may be even possible to make a profit. It should be noted that vitrification is the only method that allows such a transformation of the inert materials produced.

In terms of experimental research (especially in Italy, Britain, and Spain), over the past few years cellular glass for thermal and acoustic insulation, glass-ceramic materials for building applications (such as paving tiles), insulating or reinforcing vitreous fibers, glass, or glass-ceramic matrix composites were obtained, while just a few plants for the production of insulating fibers have been so far realized at an industrial level.

Treatment of High-Level Radioactive Waste

In the 1960s, it was recognized that the immobilization of high-level radioactive waste (HLW) in glass was the most appropriate treatment to obtain an extremely durable and long-term stable waste form. Significant advantages of the glass waste form include, in fact, its tolerance in terms of variability in chemical composition and its processability. At present, vitrification is internationally still accepted as the best demonstrated available technology [1, 2]. Since the combination of high temperatures and high-level radioactivity in the course of the vitrification process requires particularly sophisticated technologies to meet all safety requirements, a large number of experiences and tests was carried out in the past. The vitrification plant of Marcoule (France) was the first one starting worldwide to operate with HLW originating from spent fuel reprocessing plants (1978) [13]. The process was carried out in two stages: in the first one, successive steps of evaporation, drying, and calcination of the waste and the addition of borosilicate glass frit slurry were performed in a rotary tube; in the second one, the calcined product was melted – in a discontinuous process – inside a metallic pot heated by a multi-zone induction furnace. The melt, after homogenizing, was drained into a stainless steel canister, where it solidified transforming into an inert and stable glass form. The canister was then sealed and decontamined to be isolated in a deep geological formation. Emissions generated in the melter from the vitrification process were treated by an off-gas system to remove radioactive contamination and destroy nitrogen oxides (NOx). This technology was adopted by the vitrification plants in La Hague (France) and, with a few modifications, in Sellafield (UK), in the 1980s.

Because of the limited throughput and short melter life – the latter due to high temperatures glass corrosion – a more advanced vitrification system was then developed, starting from the second half of the 1970s [14] based on a Joule-heated ceramic melter (JHCM). In the JHCM, thermal energy is generated by passing an electric current through the molten glass across multiple pairs of electrodes. Since glass in the molten state is an ionic conductor of electricity, it can increase its temperature due to Joule effect. This technology was developed mainly in the USA [15] and, afterward, in Germany [16]. Operational plants were built in Belgium (Mol), the USA (Savannah River, West Valley), and Japan (Tokai Mura). At present, JHCM technology is the standard vitrification method for high-activity waste worldwide. Advantages of the JHCM over metallic pot induction heated melters are: capability for direct liquid feeding, higher throughput, enhanced life of the melter (whose walls are made of high corrosion resistant alumina-zirconia-silica refractory), and better operational flexibility, due to an unrestrained heat transfer area and the possibility of adopting a continuous mode of operation. Limitations of the process are: shortened life of the electrodes (if they

are utilized at temperatures in excess of 1,200°C), possibility of electrical short-circuiting (causing failures in the melter) due to redox conditions and presence of free metal, and difficulties in the decommissioning of the ceramic melter at the end of its life.

In the last few years, a new attractive technology has been emerging, that is, cold crucible induction melting (CCIM) [17]. This process is a modification of induction heating in which cooling is applied to form a protective layer of solidified glass on the internal walls of the crucible. This technology looks very promising for the future, and could be applied to many other types of hazardous waste, in addition to radioactive waste.

Vitrification Processes and Plants

The various vitrification processes can be distinguished according to the heating method employed in the furnace, into which the waste to be vitrified are continuously fed. High temperatures are generally used, and often the heterogeneous nature of the waste may give rise, in addition to the vitrified material itself, to other products that must be removed safely. For example, molten metal can be deposited at the bottom of the furnace, while at the surface a thin layer of liquid salts can form, especially if the batch contains a significant amount of chlorides, sulfates, or other species of limited solubility in molten glass. The output stream of gas and vapors is comprised (besides of the combustion products) of organic substances, often found in the batch, or their pyrolysis products, the most volatile heavy metals (such as mercury and cesium), particularly if reduced by carbon or other reducing substances contained in the feed, heavy metals, chlorides, particulate (dust), etc. These substances must be properly separated and collected, or chemically treated, to be recycled in the process or disposed of separately in a landfill. Thus, it can be inferred that the choice of a vitrification process cannot be limited to a simple choice of the melter, but should also consider systems and installations for the treatment of secondary effluent.

The main types of melters used for the vitrification of waste are:

Furnaces based on conventional combustion [18–22]. The thermal energy necessary for waste melting is provided by the combustion of natural gas, fuel oil, or coke. These melters can be of various types:

- *Tank melters*, such as those used in the production of flat or container glass, but with some modifications. The waste feed is introduced onto the glass pool, which is contained in a refractory-lined tank. Burners are directed at the waste feed to provide the required energy. These melters are characterized by high reliability, long life, ability to vitrify waste of very different composition.
- *Cyclone melters*, in which the particulate waste feed is incorporated in the flowing combustion gases which are forced to spiral around the inside of the melter. In this way, the particles are melted and the melt runs down the walls of the cyclone toward an orifice in the bottom. Although the high processing rate is advantageous, the control of composition and other processing parameters may be difficult. These melters, similarly to rotary furnaces, are often preferred in the presence of a high content of organic matter, but can give an inhomogeneous melt, requiring an additional conditioning unit.
- *Surface melting furnaces*, in which the waste are pushed onto the refractory-lined furnace floor, whose inclined surface receives radiation heat from burners located on the roof.
- *Small-scale blast furnaces*, which use coke for burning, but the highly reducing processing conditions increase the loss of heavy metals, chlorides, and sulfur in the form of gaseous effluents.

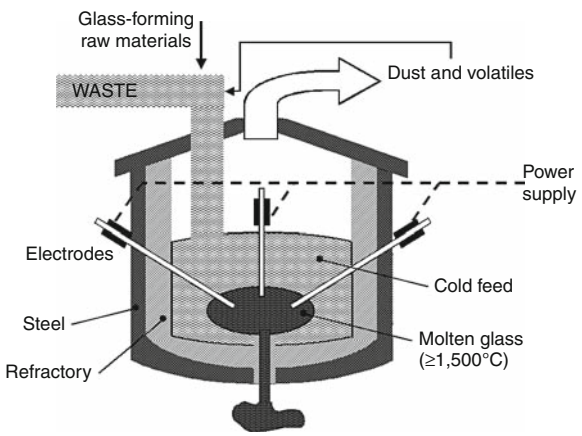
Ultimately, the disadvantages of all these melters consist especially in the volatilization of heavy metals, in particulate-rich exhaust gases and in the large volume of off-gases, which must be purified in high cost facilities. The use of oxygen instead of air as oxidizer increases the direct operating cost, but decreases by over 70% the volume of effluent gases, greatly reducing the capital and operating costs of the purification systems.

Fluidized bed gasifying melting systems [23]. A new technology for the incineration of waste has been developed in Japan [24]. The waste, milled and dried if necessary, are treated (at temperatures between 500°C and 650°C) in a fluidized bed gasifying plant, with the generation of a gas stream and a solid residue. Both of these products are then completely oxidized (at temperatures between 1,250°C and 1,450°C) in a two-stage vortex combustion plant, in which dioxins are completely destroyed and the ashes are melted to

obtain a vitreous slag. The gas enthalpy is recovered in a particular heat exchanger with circulation of alumina grains, which enables preheating of the combustion air without problems of fouling or corrosion. The gaseous effluents are then cooled, de-pulverized by bag filter, and treated by a NOx destruction system.

Electric furnaces heated by Joule effect [5, 18, 19, 21, 22]. In this type of melter, an electric current is passed by ionic transport through the molten charge from one electrode to another between which a potential difference is applied. The electrodes (one or two couples) are positioned in the hearth or in the side walls of the melter and are immersed in the melt. The electric energy of the current is dissipated as heat by Joule effect, with a very high conversion efficiency. Since the electrical conductivity of the initial solid charge is negligible and starts to have an appreciable value only above ~800°C (when the heated material gradually melts and becomes an electrolytic conductor), the charge must first be melted by a secondary heating system (e.g., by resistors or by special burners). Once the initial charge is completely melted, further batch is fed continuously from the top of the melter and, falling into the melt, is in turn heated and fused at expense of the available energy, while in parallel a certain amount of molten glass is discharged from the melter. The molten material is more or less conductive depending on its chemical composition, and this should be taken into account for setting the temperature of the process, since electric conductivity increases with temperature. The lack of combustion gases entails a drastic reduction of the gaseous effluents; their volume may in fact be even two orders of magnitude smaller than for combustion-based melters, so off-gas treatment plants (needed in the presence of hazardous organic materials) are of minimal size and rather simple. This type of melter is particularly suited to process homogeneous inorganic materials that are finely divided, such as sludge, soils, and crushed concrete. To increase the efficiency of the process, the feedstock must be accurately controlled; if it is not finely divided, it is slow to dissolve and, besides, inappropriate materials can slow down the melting or even corrode the electrodes. The molten material, once extracted from the melter and cooled, solidifies in a glassy form. To achieve the waste vitrification, that is, its complete transformation into a homogeneous and chemically durable glass, vitrifying additives (to be chosen among other waste, e.g., cullet) may be added and intimately mixed to the charge before it is fed to the melter.

Joule-heated melters have a refractory-lined cylindrical chamber made of stainless steel and cooled by a water jacket. An opening on the bottom allows for the discharge of molten material through a channel. The crown, that is also refractory-lined, has an opening for the introduction of the feedstock and one for the off-gas release (see Fig. 1). To prevent the volatilization of heavy metals, cold top Joule-heated melters are used, that is, with the melt surface kept at low temperature, because it is covered by a layer of feedstock continuously supplied and not yet melted. In this case, the volatile substances released at the bottom of the furnace, which is warmer, ascend and condense upon contact with the cold top, and are then mixed again with the residual melt by convection currents. Even salts (as chlorides or sulfates) condense, forming a liquid layer at the interface between feedstock and melt, and can be removed with special drainage techniques. The cold blanket acts as a thermal insulation shielding the crown from thermal radiation; the crown thus remain at low temperature and the off-gas is cooled (to about 150°C). Additional heating elements at the top of the melter (hot top Joule-heated melters) can be inserted to obtain the complete destruction of hard to treat hazardous organics, but in this case there is an increase of particulate and volatile species in the



Vitrification of Waste and Reuse of Waste-Derived Glass. Figure 1
Sketch of an electric furnace heated by Joule effect

off-gas. Intensive stirring inside small volume Joule-heated melters results in high throughput. The electrodes, usually made of molybdenum or graphite, are corroded by free metals if they are contained in significant amounts in the waste, so that very small quantities of unoxidized metals can be tolerated. In some cases, more expensive electrodes (made of tin oxide or chromium oxide) can be utilized. The type of electrode used may strongly influence the maximum operating temperature. Moreover, the free metals can alter the electric current distribution within the melt and cause a significant decrease in the efficiency of the process. Both water and organic substances in the feedstock should never exceed 5 wt%, as they require large amounts of energy for heating, evaporation, gasification and molecular dissociation, as well as an increase in the cost of the plant because of the need of an off-gas treatment equipment. In any case, to ensure process reliability parameters, a careful formulation of the batch composition is required.

Terra-Vit melter [25, 26]. The Terra-Vit melter was designed to increase the life of the melter and reduce capital and operating costs. It is a Joule-heated melter and consists basically in a pit of semi-spherical form, that is excavated in the ground and covered by a refractory brick roof. Electrodes are introduced into the molten pool through holes in the roof; their position must be such as to limit corrosion and erosion of side walls and floor, which are lined with refractory clays and can be water-cooled to further limit corrosion. Waste is continuously fed through the roof onto the molten pool and melts progressively. The molten material is discharged through an opening in the wall, located at the same level of the melt surface. If the waste contains organic or combustible fractions, oxidation air is blown onto the molten pool surface. It is also possible to add some coal to the waste to burn these components and therefore lower the energy costs. If the waste feed contains steel or other metals, they melt and accumulate on the bottom of the melter, and must be oxidized with injected air.

Electric arc furnaces [18–22]. This technology has been adopted from steelmaking, and is most applicable to mainly inorganic, dry waste. The melting of the feedstock is caused by the heat produced by means of the arc generated by an electrical three-phase current passing through three graphite electrodes, which penetrate into the furnace (see Fig. 1). The previously dried solid charge is fed from the top of the furnace and falls, spreading out, on the already melted material. The furnace operates at high temperature (>1,400°C), and therefore any ferrous materials contained in the charge can be melted evenly within a short time. The slag exceeding a given level is continuously released from the melting tank through a refractory channel, in which an additional electrode can be inserted. The melted slag is then quenched with water and carried out by conveyor. Any organic components in the charge are burnt completely in this type of furnace, and are then removed by the exhaust gas. The atmosphere within the furnace is maintained oxidizing and there are no water-cooled parts; therefore, risks of explosion are minimized. These furnaces are characterized by structural simplicity, low thermal losses, and high throughput.

Plasma torch furnaces [27–30]. The process is similar to that which occurs in electric arc furnaces, except for the use of a thermal plasma to convey the arc energy. The electric arc is generated by a potential difference applied between two electrodes, and may be either of a non-transferred (both the metallic electrodes are confined within the torch) or a transferred type (from one electrode inside the torch to another of graphite inside the hearth). The electric arc overheats a processing gas (air in case of waste) flowing inside the torch, causing its ionization. The thermal plasma thus obtained may reach very high temperature (even higher than 15,000°C), and heat transfer occurs mostly by radiation. The plasma jet hits with high energy density a limited area of the waste, whose temperature is raised to 3,000–4,000°C. The organic substances contained in the waste undergo sublimation, molecular scission, reforming reactions, and carbon oxidation by steam injected in the kiln in a controlled amount (e.g., $C + H_2O \rightarrow CO + H_2$), with complete gasification and production of a synthetic gas named "syngas." Syngas is a fuel gas which is composed primarily of H_2 and CO, with varying percentages of N_2, CO_2, water vapor, CH_4, and other light hydrocarbons (depending on the chemical composition of the feedstock). It undergoes various cleaning treatments aimed at (a) removing particulate matter (which is then collected and batch-fed back into the plasma vessel), (b) hindering the formation of dioxins and furans by quenching from

high temperatures, (c) performing selective catalytic reduction of NOx, and (d) dissolving any acid gas and other inorganic ions (chloride, fluoride, sulfate, phosphate, sodium, and calcium) into the liquid of a packed column scrubber. After purification, syngas can be used as fuel for electricity generation, or for the production of precursors (e.g., methanol, ethanol) for the chemical industry or for the separation of pure hydrogen through ultrafiltration. The hydrogen from syngas can find use in the petrochemical (hydrogenation of petrol) or food (hydrogenation of fats) industry, and in hydrogen fuel cells which can power transportation (cars, trucks, buses, and ships).

The inorganic fraction of waste treated by plasma is melted and transformed, after being extracted from furnace and cooled, into a vitrified material similar to a glassy volcanic rock such as obsidian, in whose matrix the heavy metals (together with any contaminant from the syngas purification sections) are incorporated and completely inertized. The obtained material, because of its very low leachability, can be utilized without environmental risks as secondary raw material for road embankments, concrete mix, filling, etc. The plasma vessel is a cylindrical stainless steel container, lined with refractory materials, that is maintained at a slight negative pressure to ensure that no gases can escape to atmosphere. The plasma torch is inserted in the crown (see Fig. 2a); there are openings for the introduction of solid, liquid, and gaseous feedstock, by themselves or in any proportion or combination, and an exit port in the bottom part to remove excess molten material (followed by cooled rollers – see Fig. 2b – or not, with the direct casting in metal containers – see Fig. 2c).

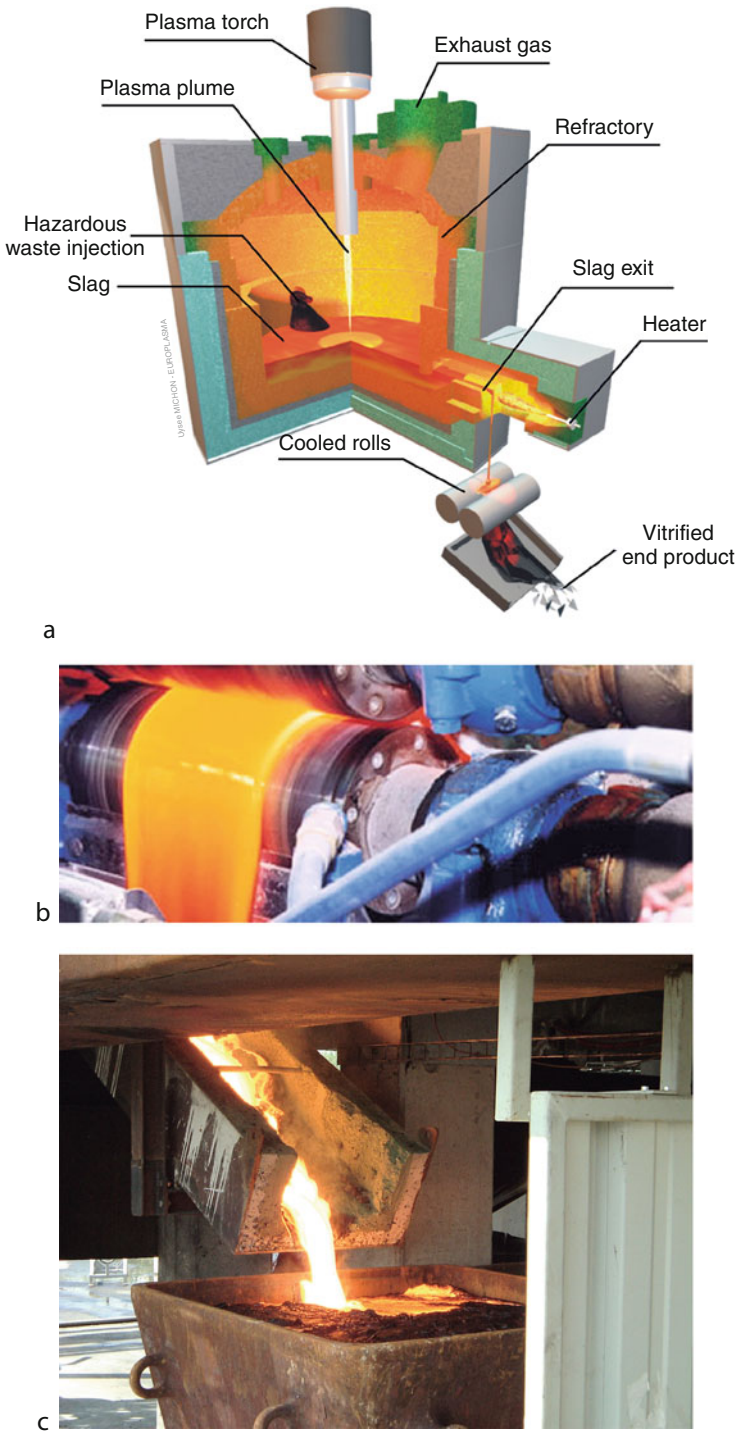
The advantages of this technology are:

- Very wide flexibility (it is possible to treat every kind of waste materials – even those containing large amounts of metals – in any physical state and with large variations in composition and ratio between organic and inorganic components).
- The power input can be quickly adjusted to match process requirements.
- The process can be stopped and restarted easily, eliminating the need for large waste storage.
- High efficiency – because of very high temperatures – of destruction of organic compounds (including hazardous organic micropollutants such as dioxins, furans, and polycyclic aromatic hydrocarbons).
- Lack of waste products (particulate and sewage sludge from the cleaning treatments of the syngas are sent back to the reactor to be vitrified).
- Compact plant (and consequent possibility of realization of mobile units).

The drawbacks of this technology are: limited yield of the plant, very expensive technology because of the high energy consumption (much higher than for other types of furnaces) and frequent maintenance operations (duration of the torches is very limited, because of the consumption of electrodes). Moreover, the maximum water content in the waste to be treated must be less than 10 wt% and the net heat value of the waste must be at least 12–15 MJ/kg, otherwise the whole process becomes uneconomical.

Because of these reasons, plasma technology is particularly suited for the treatment of hazardous waste which require very high temperatures to be totally destroyed (e.g., concentrated organic waste types, including PCBs, pesticides, and persistent organic pollutants) and therefore cannot be eliminated by using technologies such as incineration, gasification, or pyrolysis, or need operating conditions that can minimize the risk to human health and the environment (e.g., asbestos containing waste, fly ash from incineration of MSW, hospital waste, etc.).

Cold-crucible induction melters [17, 31, 32]. This technology has been so far mainly applied to the vitrification of highly radioactive waste. The metallic crucible of an induction heated melter is cooled in order to cause the formation, on its inner walls, of a solidified glass layer, which protects the crucible against glass corrosion. The inductive coupling of a water-cooled high frequency electrical coil with the glass causes eddy currents with the production of heat and effective stirring of the melt. Since the molten glass is contained inside a frozen layer of the same composition, and because of the lack of electrodes, no contamination can occur and melts with high purity are attainable. As there are no limitations arising from the corrosion of refractories or failure of the electrodes, high process temperature (up to 3,000°C) can be achieved for the treatment of waste otherwise difficult to process.



Vitrification of Waste and Reuse of Waste-Derived Glass. Figure 2

(**a**) Example of a plasma torch furnace; (**b**) vitrified MSW incinerator fly ash cooled between rollers; (**c**) pouring of vitrified asbestos-containing waste (Courtesy of Dr. Erika Edme, Europlasma, Bruges, France – http://www.europlasma.com)

Moreover, the reaching of high temperatures increases the processing rate and allows for greater throughput.

In addition to employing variously heated melters, two other vitrification technologies are currently in use, namely direct in situ vitrification and self-sustaining vitrification.

In situ vitrification [7, 8]. In situ vitrification (ISV) is a waste treatment technology that uses electrical power to melt in place contaminated earthen media such as soil, sediments, and mine tailings. It has been marketed by GeoMelt^{TM} (Richland, WA-USA) from the early 1990s and has been applied the first time at a Superfund site in Grand Ledge, Michigan (1993–1994) to treat soils and sediments contaminated with pesticides, metals, and dioxins. Afterward, several contaminated soils and debris have been successfully remediated in the USA, Japan, and Australia. This type of vitrification technology consists of four electrodes (placed symmetrically and fed by a two-phase alternating current), which are immersed into a portion of the contaminated soil circumscribed by the electrodes and melt it completely, causing its vitrification into a monolithic block. The solid ground, poorly and only occasionally conductive, is initially heated to melting by inserting between the electrodes (to which a potential difference is applied) a conductive medium, formed, for example, by a strand of glass fibers, in itself highly insulating, impregnated by electrically conductive graphite. When power is supplied to the electrodes, this medium conducts the current through the soil, heating the surrounding area and forming a pool of molten soil at the surface of a treatment zone between and directly adjacent to the four electrodes. The electrically conductive molten pool serves then as the heating element for the process, wherein electrical energy is transformed into heat by Joule effect. As the soil becomes molten, the electrodes are lowered further into the soil, continuing the melting process to the desired depth of treatment. In this way, the molten mass grows outward and downward (up to a maximum area of ~12 m in diameter and to a depth of 6–7 m), thereby creating individual batch melts at an average rate of approximately 3–4 tons per h. Melt temperatures are in the range of 1,500–2,000°C, and energy consumption varies from 500 to 800 kWh/ton in most applications, so that the process is very efficient. When the desired volume has been melted, power is shut off and the molten mass cools to form a hard, monolithic product of glass and crystalline material that incorporates and immobilizes the inorganic compounds and metals. This material is stable and chemically inert, with very low leaching characteristics, and is effectively and permanently isolated from the surrounding environment. The high temperature of the melt promotes pyrolysis and/or vaporization of all the organic substances in the hot soil surrounding the melt. Consequently, organic contaminants and any other volatile constituent escape and migrate to the surface. In the GeoMelt^{TM} system, a dome-shaped hood completely covers the treatment area, supports the electrodes and collects the gaseous emissions, which are then conveyed to the wet and dry off-gas treatment system. Combined with the destructive action of the melt, the off-gas treatment system contributes with very high efficiency to the further destruction and/or removal of trace quantities of contaminants present in the off-gas. The ISV equipment is mobile and can be repositioned in another area of the contaminated soil after any treatment. It is also possible to collect, melt, and vitrify waste of different types, or coming from different locations, that are enclosed in one or more reusable cells. Cost is affected by the depth of the contamination, the degree of homogeneity of soil, the presence of debris, and excess moisture. Moreover, in several countries the price of electricity can be an important factor. The generally high cost of the process makes it especially suited for hard to treat hazardous wastes, such as mixtures of metals and organics. The process is highly effective for remediating soils contaminated with heavy metals (including Cr, Pb, Cd, and As) and virtually all chlorinated organic compounds (including solvents, pesticides, herbicides, dioxins, furans, and poly-chlorinated biphenyls). Even sites contaminated by radioactive materials have been successfully remediated using the GeoMelt^{TM} process.

Self-sustaining vitrification [10, 33]. Self-sustaining vitrification uses the energy released during exothermic chemical reactions (similar to the well-known thermite reactions) that occur in a mixture of waste and powder metal fuel (PMF) to form a melt, which cools down to produce a glass-like material. The process is controlled by the composition of the initial mixture of dried and crushed waste and PMF; it does not require either an external power supply or large and expensive

equipment, and it is economically viable particularly for small volume hazardous waste. PMF is a specially designed blend of heat generating components; generally, it consists of combustible powder metals (as Al, Mg), oxygen-containing components, and some additives such as stabilizers and surface active substances. A suitable PMF composition must: (1) release an amount of heat sufficient to sustain waste melting without external heating, (2) produce a glass-like end product binding contaminants in its structure, and (3) minimize carryover of toxic chemicals, heavy metal vapors, oxides, and radionuclides. After initial ignition, a combustion wave propagates within the mixture at a velocity of ~1 mm/s; PMF must therefore be tailored in order to maintain the reaction parameters within a controlled and stable range, providing a good quality product. Self-sustaining vitrification has been demonstrated to be a feasible immobilization technology for many types of waste, including radioactive waste, [33] contaminated soils [34] and ashes [34].

The vitrification technologies most suited to the treatment of waste are reported in Table 1, along with their main characteristics.

To summarize, several technologies are currently available for the vitrification of waste, and the choice of which one to adopt depends on several parameters, among which are: the composition, form, amount and geographical location of the waste to be treated, the cost of the plant itself (including treatment of off-gases), the cost of energy, the cost of labor, the flexibility in terms of waste composition accepted (a single plant could serve several industries, but then the cost of waste transport should be added), the issuing of economical incentives from local government agencies, the existence of local laws regulating the handling/recycling of special waste, etc. Therefore, the choice of the most suitable technology, even for the same typology of waste, will vary with the location of the plant and the local political and labor circumstances.

Waste Sources

A definition of hazardous substance or hazardous waste is neither easy nor simple. A legal definition of a hazardous waste by EPA (the US Environmental Protection Agency) is: (1) any waste or combination of waste of a solid, liquid, contained gaseous, or semisolid form which cause or significantly contribute to an increase in mortality or an increase in serious irreversible or incapacitating reversible illness, or (2) pose a substantial present or potential hazard to human health or the environment when improperly treated, stored, transported or disposed of, or otherwise managed [35]. EPA compiled a detailed list of hazardous materials that includes tens of thousands of chemicals, and pointed out six tests that enable to classify a substance as hazardous. These tests measure specific properties of the material, such as: (1) level of radioactivity, (2) bioconcentration, (3) flammability, (4) reactivity, (5) toxicity, (6) genetic, carcinogenic, mutagenic, and teratogenic potential. According to EPA, if a substance does not fail any of the six tests, it is not hazardous, and it is possible to dispose of it in ordinary landfills.

In the European Union, the List of Waste (formerly the European Waste Catalogue), is a catalogue of all waste types generated in the EU. It was established by Commission Decision 2000/532/EC of May 3, 2000 [36] and has been amended by Commission Decision 2001/118/EU, 2001/119/EU, and 2001/573EU. The List of Waste classifies all waste, whether intended to disposal or recovery, in 20 chapters (see Table 2), according to the activity that originated them. Each of the 20 chapters is grouped into subchapters and includes a detailed list of specific waste. Every waste is fully defined by a six-digit code, with two digits each for chapter, subchapter, and waste type. Hazardous wastes are those marked with an asterisk next to their appropriate identification code.

The European List of Waste is a harmonized list of waste and is subject to regular revisions and/or amendments, to take into account scientific and technical progress. Waste classified as hazardous are considered to display one or more of the properties listed in Annex III to Directive 91\689\EEC (Table 3) and, in reference to labels H3–H8, H10, and H11 of the said Annex, one or more characteristics related to their flash point or to the concentration limits of some specific pollutant contained in the waste (Article 2 of the Commission Decision 2000/532/EC, as amended).

There are test methods which serve the purpose to give specific meaning to the definitions reported in Annex III. The methods to be used are those described in Annex V to Directive 67/548/EEC, in the version as

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 1 Main technologies used to vitrify waste

Vitrification technology	Heat source	Advantages (A), disadvantages (D), and limitations (L)
Combustion based melters (tank melters, cyclone melters, surface melting furnaces and small-scale blast furnaces)	Combustion (natural gas, fuel oil, or coke)	(A) High reliability, long life, ability to vitrify waste of very different composition (tank melters). High processing rate (cyclone melters) (D) High capital cost. Inhomogeneous melt, volatilization of heavy metals, particulate in the off-gas, large volume of the off-gas (to be treated in large and expensive depuration plants). Pure oxygen as combustion agent decreases the off-gas volume by over 70% (but extra-cost has to be considered)
Fluidized bed gasifying melting systems	Combustion (natural gas, fuel oil, or coke)	(A) Optimum thermal efficiency. Complete destruction of toxic contaminants as dioxins. Melting of ashes into a vitreous slag
Joule-heated melters (JHMs): cold top JHMs, hot top JHMs, intensively stirred small volume-high throughput JHMs	Electrical power (Joule heating, due to ionic conduction inside the melt)	(A) Drastic reduction of the off-gas volume and thus of the purification plant size. By employing cold top JHMs it is possible to hinder volatilization of heavy metals and to obtain salts condensation. With hot top JHMs complete destruction of hazardous organics is obtained (D) Free metals attack graphite or Mo electrodes and decrease the process efficiency (L) System particularly suitable to vitrifying properly divided homogeneous inorganic materials (mud, soils, and finely ground concrete). Careful formulation of the charge is required: water and organics must not exceed 5 wt%
Terra-Vit melter (JHM)	Electrical power	(A) Reduced capital and operating costs. Increased melter life
Electric arc furnaces	Electrical power (electric arc generated by three electrodes on the furnace top)	(A) Systems characterized by simplicity of construction, low thermal losses, high output. High process temperature ($>$1,400°C). Fast ferrous materials melting. Risks of explosion minimized (L) Technology most applicable to essentially inorganic, dry waste
Plasma torch melters	Electrical power (ionization of a process gas by the electric arc generated, and formation of a very high temperature plasma)	(A) Compact plant (mobile units can be used). Very high temperatures attainable. High efficiency of toxic organic compounds destruction. Very wide flexibility, that is, possibility of direct treatment of highly different types of waste, even containing large quantities of metals. Absence of waste products (D) Limited yield of the plant. Plasma torch melters consume more energy, need frequent maintenance and are less durable than combustion-based or electric furnaces (L) Technology most suited to waste requiring high destruction temperature or operating conditions that minimize risks to human health and the environment
Cold-crucible induction melters	Electrical power (induction heating of a water-cooled crucible)	(A) Crucible protected against corrosion. Effective stirring of the melt. High purity melts attainable. Very high processing temperatures achievable, if required
In situ vitrification	Electrical power (Joule heating of contaminated soils)	(A) Mobile equipment. High melting temperatures attainable. Highly effective process for remediating soils contaminated by heavy metals, chlorinated organic compounds and radioactive materials (L) High cost of the process makes it sustainable only for hard to treat hazardous waste

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 1 (Continued)

Vitrification technology	Heat source	Advantages (A), disadvantages (D), and limitations (L)
Self-sustaining vitrification	Exothermic chemical reactions between powder metal fuel and the waste	(A) No need of external power supply. Large and expensive equipment not required (L) Technology most suited for small volume hazardous waste

amended by Commission Directive 84/449/EEC13, or by subsequent Commission Directives adapting Directive 67/548/EEC to technical progress. These methods are themselves based on the work and recommendations of the competent international bodies. Selected examples of important and representative categories of hazardous waste are reported in Table 4.

Vitrification has been investigated and successfully achieved for many of the waste listed in Table 4, and reports on the operation of vitrification plants can be found in the literature [37].

It should be stressed that the chief aim of vitrification is to produce a homogeneous glass possessing a high chemical durability, and detailed and complete tests have to be performed in order to be able to predict the long-term behavior of the material under the complex conditions prevailing in a repository environment or in an actual reuse of the glass. Because of the importance of this aspect, some studies tried to understand and predict the correlation between leaching behavior and composition of the waste or of the glasses [38, 39]. Furthermore, thermochemical modeling of oxide glasses or the determination of solubility limits of some pollutants in oxide glasses allows to predict crystal formation or phase separation, both greatly affecting leaching behavior [40, 41]. Some authors assessed the risk posed by leachates coming from waste-derived glasses (obtained from fly ash), proving indeed that they pose a minimum ecotoxicological risk, thus opening the possibility of reusing such glass for the development of commercial products [42].

The most frequently treated waste typologies will be briefly discussed below [43].

Slags from ferrous and nonferrous metallurgical industries are the first industrial waste that were vitrified with the aim of inertization and recovery producing, after crystallization, tiles possessing properties suitable for the building sector. The first example dates back to the early 1960s [44] followed by other examples, employing blast-furnace slag, which were reported in the early literature [45–47]. They are produced in large volume and are characterized by a high content in glass-forming and network-modifying oxides (SiO_2, Al_2O_3, CaO). In most cases, vitrification is followed by crystallization, either during cooling or via a controlled secondary heat treatment, and the process is facilitated by the presence of large amounts of elements prone to the formation of crystalline phases (such as iron oxides) or by the addition of suitable nucleating agents (such as TiO_2, Cr_2O_3, P_2O_5, sulfides, and fluorides) [44]. While blast-furnace (BF) slag is widely used in the cement industry, the vitrification route is at present the only treatment suitable for the recovery of basic oxygen furnace (BOF) slag (up to 60%, added to sand and Na_2O), because of its higher iron and lower silica content [48]. Compositional adjustment have been required in some cases to form a stable glass upon cooling. For instance, blast-furnace (BF) slag was added to ash from a coal-fired electric power station and a waste product from the copper producing industry, producing a homogeneous glass, which was subsequently crystallized [49]. Steelwork slag has also been successfully vitrified, [50] with the addition of residues of bauxite extraction, limestone, sand, and TiO_2 as nucleating agent (added as ilmenite), forming then an iron rich glass-ceramic. Slag deriving from the hydrometallurgy of zinc ores, containing a large amount of Fe (~50 wt%), of Zn (~6 wt%) and of Pb (~4 wt%), has been also vitrified and ceramized (with the addition of 30–60% of cullet), producing materials with a Fe content higher than 20 wt% [51, 52]. The durability depends on the

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 2 Chapters of the List of Waste (EU)

01.	Waste resulting from exploration, mining, dressing and further treatment of minerals and quarry
02.	Waste from agricultural, horticultural, hunting, fishing, and aquacultural primary production, food preparation and processing
03.	Waste from wood processing and the production of paper, cardboard, pulp, panels, and furniture
04.	Waste from the leather, fur, and textile industries
05.	Waste from petroleum refining, natural gas purification, and pyrolytic treatment of coal
06.	Waste from inorganic chemical processes
07.	Waste from organic chemical processes
08.	Waste from the manufacture; formulation; supply; and use (MFSU) of coatings (paints, varnishes, and vitreous enamels), adhesives, sealants, and printing inks
09.	Waste from the photographic industry
10.	Inorganic waste from thermal processes
11.	Inorganic metal-containing waste from metal treatment and the coating of metals and nonferrous hydrometallurgy
12.	Waste from shaping and surface treatment of metals and plastics
13.	Oil waste (except edible oils, 05 and 12)
14.	Waste from organic substances used as solvents (except 07 and 08)
15.	Waste packaging; absorbents, wiping cloths, filter materials, and protective clothing not otherwise specified
16.	Waste not otherwise specified in the list
17.	Construction and demolition waste (including road construction)
18.	Waste from human or animal health care and/or related research (except kitchen and restaurant waste not arising from immediate health care)
19.	Waste from waste treatment facilities, off-site waste water treatment plants and the water industry
20.	Municipal waste and similar commercial, industrial, and institutional waste including separately collected fractions

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 3 Properties that render waste hazardous (EU)

H1"Explosive": substances and preparations which may explode under the effect of flame or which are more sensitive to shocks or friction than dinitrobenzene
H2 "Oxidizing": substances and preparations, which exhibit highly exothermic reactions when in contact with other substances, particularly flammable substances
H3-A "Highly flammable": - liquid substances and preparations having a flash point below 21°C (including extremely flammable liquids), or - substances and preparations which may become hot and finally catch fire in contact with air at ambient temperature without any application of energy, or - solid substances and preparations which may readily catch fire after brief contact with a source of ignition and which continue to burn or to be consumed after removal of the source of ignition, or - gaseous substances and preparations which are flammable in air at normal pressure, or - substances and preparations which, in contact with water or damp air, evolve highly flammable gases in dangerous quantities
H3-B "Flammable": liquid substances and preparations having a flash point equal to or greater than 21°C and less than or equal to 55°C
H4 "Irritant": noncorrosive substances and preparations which, through immediate, prolonged or repeated contact with the skin or mucous membrane, can cause inflammation
H5 "Harmful": substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may involve limited health risks
H6 "Toxic": substances and preparations (including very toxic substances and preparations) which, if they are inhaled or ingested or if they penetrate the skin, may involve serious, acute or chronic health risks and even death
H7 "Carcinogenic": substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may induce cancer or increase its incidence
H8 "Corrosive": substances and preparations which may destroy living tissue on contacts

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 3 (Continued)

H9 "Infectious": substances containing viable microorganisms or their toxins which are known or reliably believed to cause disease in man or other living organisms
H10 "Teratogenic"9: substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may induce nonhereditary congenital malformations or increase their incidence
H11 "Mutagenic": substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may induce hereditary genetic defects or increase their incidence
H12 Substances and preparations which release toxic or very toxic gases in contact with water, air, or an acid
H13 Substances and preparations capable by any means, after disposal, of yielding another substance, for example, a leachate, which possesses any of the characteristics listed above
H14 "Ecotoxic": substances and preparations which present or may present immediate or delayed risks for one or more sectors of the environment

crystalline phase assemblage, and can be higher than that of the parent glass [53]. Finally, waste coming from a geothermic plant and containing a large amount SiO_2 were used as silica source for the production of glass-ceramics or optical glass [54, 55].

Fly ash, which derives from various industrial processes, is also one of the most important examples of waste that have been successfully vitrified by several researchers [56]. In particular, fly ash deriving from municipal solid waste (MSW) incineration is one of the most extensively investigated residues. In fact, early studies demonstrated that vitrification of MSW fly ash enables the immobilization of hazardous metals and destroys the organic pollutants, [57, 58] with the advantage, at the same time, of achieving a significant reduction in the volume of waste (as high as 80–90%). Such waste consists of the finer fraction of ash separated from the effluent gas by the filters, and represents a serious environmental problem because it contains a significant amount of hazardous organic and inorganic constituents, such as dioxin, furans, and heavy metals (mainly Cd, Cr, Cu, Pb). Typically, the content of glass formers (mainly silica) in this type of waste is

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 4 Selected examples (EU) of hazardous waste (HW)

HW from the manufacture, formulation, supply, and use (MFSU) of acids; from the MFSU of halogens; from the MFSU of basic organic chemicals; from the MFSU of plastics, synthetic rubber and man-made fibers; from the MFSU of organic solvents
Bottom and fly ashes (e.g., from incineration of MSW) containing dangerous substances
Waste from the iron and steel industry containing dangerous substances
HW from Al, Pb, Cu, Zn, and other nonferrous thermal metallurgy
HW from casting of ferrous and nonferrous pieces
HW from petroleum refining
HW from the MFSU of pharmaceuticals; from the MFSU of fats, grease, detergents, soaps, disinfectants, and cosmetics; from the MFSU of fine chemicals
HW from the MFSU and removal of paint and varnish; from the MFSU of adhesives and sealants
HW from the photographic industry
HW from electrical and electronic equipments
Insulation materials containing asbestos
Lead batteries, Ni-Cd batteries, mercury-containing batteries
Waste organic solvents, refrigerants, and propellants
Oil waste and waste of liquid fuels
HW from galvanic processes, zinc coating processes, pickling processes, etching, phosphating, alkaline degreasing, anodizing
Waste explosives
MSW hazardous components (fluorescent tubes; batteries and accumulators; adhesive, paint, inks; discarded electrical and electronic equipments; detergents and solvents)
Hospital and medical waste
Soil, stones, and dredging spoil containing dangerous substances

rather low (<35 wt%), and therefore the addition of other raw materials (e.g., cullet or silica) to correct the composition of the batch is needed. However, several examples of direct vitrification of MSW fly ash without

any cullet addition exist, [59–61] and some reports show the successful formation of a (highly crystallizable) glass starting from fly ash containing <12 wt% of silica [62, 63]. Due to the low content of network formers, the durability of the glasses resulting from the vitrification of fly ashes needs to be carefully evaluated, to verify the lack of heavy metal leaching [64] and the degree of inertness of the material. In some cases, the presence of salt (added to food) and plastic packaging in the ash residues constitutes a problem, as it is a source of chlorine, which is difficult to dissolve into the glass and can lead to the formation of dioxins during cooling by reaction with hydrocarbons [65]; this difficulty may however be overcome by the vitrification of fly ash after washing in water [66]. Often, the addition of alkali-containing chemicals (such as Na_2CO_3 or waste from the production of feldspar) is performed to improve the viscosity and working characteristics of the melt [67, 68]. Electric arc furnace baghouse dust (EAFD) from steelmaking processes typically contains a high amount of heavy metals (Cr, Pb, Zn, Cu, Cd, Ni), in addition to a large amount of Fe and again a low silica content, which requires the addition of glass cullet to the batch in an amount depending on the composition of the waste [69]. On the contrary, the composition of coal fly ash, which is produced in large amounts by the burning of coal in thermal power plants, is particularly rich in silica (>45 wt%) and alumina, so it can be directly vitrified, if precursors for alkaline and alkaline earth oxides are added to the batch to tailor the rheological characteristics of the melt [70]. Another type of fly ash is one rich in lead, produced by the incineration of sludge recovered from tetraethyl lead and leaded gasoline storage tanks. As they are constituted mainly of inorganic Pb compounds and iron oxide, with a silica content of ~2 wt %, a maximum amount of 60 wt% can be incorporated in a glass formed by the addition of other raw materials [67, 71].

Differently from the majority of other waste, fly ash are often not vitrified by themselves, with just the addition of minor amounts of cullet or other glass-forming raw materials. Several types of fly ash are, in fact, typically either mixed together [72–74, 76] or added as a minor constituent to a glass-forming mixture [73–76], which can itself contain other types of waste [77, 78], in order to obtain a more suitable composition of the batch. The mixing of fly ash with bottom ash may represent a further solution for the above mentioned problem of limited chloride solubility in waste glasses [65]. In all cases discussed above, the amount of fly ash that can be present in a batch depends on the general composition of the waste and the operating conditions of the vitrification process. The glasses obtained from fly ash (in limited or large amount) generally belong to the system $CaO–Al_2O_3–SiO_2$ or $CaO–MgO–Al_2O_3–SiO_2$ and, due to the frequent presence of significant amounts of components that could promote some phase separation and therefore heterogeneous nucleation (e.g., Fe_2O_3 (~10 wt%), P_2O_5 (~2 wt%), and TiO_2 (>1 wt%)), they are easily converted into glass-ceramics through a subsequent controlled heat treatment [59, 61, 67, 79–83].

Air pollution control (APC) residues, produced from the flue gas cleaning processes, represent a secondary important type of waste from incineration operations. Amutha Rani et al. [84] reported that this waste is classified as hazardous mainly due to its high alkalinity, although heavy metals, organics (primarily dioxins and furanes), and chlorides are present in significant amounts. These authors successfully applied plasma vitrification to APC residues mixed with alumina and silica.

Mud or sediments constitute another typology of waste that has been successfully vitrified. They derive either from contaminated environments (e.g., Venice lagoon dredging spoils) [85, 86] or from industrial processes (sludge from wastewater treatment, [87–89] Cr-containing mud from the tanning process, sludge from traditional ceramics cutting or polishing procedures, [78] red muds from the alumina Bayer process of bauxite or from beryllium extraction from its ore) [78, 90]. They are characterized by a large variation in their composition (especially if coming from dredging operations), the presence of a large amount of water and organic matter (frequently requiring extensive pretreatment before introduction in the melter) and often by the predominance of a specific component in their composition (thus requiring extensive compositional corrections using glass-forming materials). For a successful vitrification of this waste, a highly flexible process should be used, and a control of the average composition of the waste material before each melting campaign should be performed, to ensure the

production of a glass with chemical durability within the regulation limits. Volume reduction of the waste can be as high as 40–60%.

Among other types of waste that have also been successfully vitrified, asbestos-containing residues are of particular relevance because of the health concerns they raise. They do not contain heavy metals in significant amounts, differently from all the waste described previously; however, vitrification enables the complete destruction of the fibrous structure, which constitutes the main hazard. Because they are comprised of silicates, large quantities of such residues can be added to a glass-forming mixture, and after the thermal treatment leading to the formation of the melt, a glass of conventional composition and good chemical inertness is produced, which can be disposed of safely or reused [91, 92].

Glass cullet should also be considered in the context of vitrification of waste. Despite not being classifiable as an hazardous waste in itself, except in the case of glasses from dismantled cathode ray tubes (CRTs), featuring a remarkable content of heavy metal oxides (i.e., BaO, SrO [from the front part, known as the panel] and PbO [from the internal part, known as the funnel or cone]), [93] cullet is collected in large quantities by community glass recycling programs. Although recommended for limiting the consumption of energy and natural raw materials, the use of cullet in manufacturing traditional products (especially container glass) is possible only after an expensive sorting step, aimed at the removal of impurities (of metallic or ceramic nature). This separation leads to a fraction of almost pure glass, suitable for the industry, and a fraction enriched in contaminants, which remains practically unused, and is mostly disposed of in landfills [94, 95]. However, as mentioned before, cullet can be used to modify the composition of other waste, therefore playing the fundamental role of supplying the needed glass-forming oxides lacking from the waste to be treated. Albeit most experimental work has been carried out with soda-lime-silica glass, the Ba-Sr glass deriving from the front part of CRTs has also been used [96, 97]. Alternatively, it can be used to fabricate less conventional, glass-based products (such as glass foams), [93] or it can be added to conventional batches for the fabrication of traditional ceramic products (tiles, bricks) [98–102].

In Table 5, the various typologies of hazardous waste, for which both successful vitrification and suitable durability of the produced glass or glass-ceramic have been proven, have been grouped according to their source. It has to be noted, in fact, that the chemical-physical characteristics of waste often overlap, making it difficult to classify them according to the choice of the most valid vitrification process, from a technological and economical point of view. It is indeed because of the great flexibility of glass to accommodate large variations of composition without detrimental effects on its chemical durability, that all these experiments were successful.

In Table 6, the typical compositional range of various waste is reported, as a reference. Naturally, other waste and residues specific to particular processes and geographical locations might have widely different compositions from the ones listed below.

Despite the several successful experiences reported in the scientific literature, a fundamental difficulty in moving vitrification toward large-scale implementation consists in the fact that, typically, the incoming waste stream is largely inconsistent in composition. Naturally, this would affect greatly the day-to-day plant operations, but also make extremely difficult the reuse of the produced glass as either a vitrified product or as raw material for other industrial processes. Improvements could be achieved with more extensive and selection separation of waste, which however would require the cooperation of the waste producers as well as, in most cases, of the public.

Valorization of Waste-Derived Glasses

The vitrification process is capital and energy intensive. The process is consequently hardly sustainable, if the economic advantage is related only to the avoided disposal costs. These costs are particularly significant for highly hazardous waste, such as asbestos-containing materials, which are vitrified even by employing the most expensive technologies, such as plasma heating [105]. Even if it is rather straightforward that waste landfilling will become more and more complicated in the future, so that disposal costs are expected to rise, vitrification should be economically encouraged by favoring some sources of extra revenue, such as the use of the produced glass to fabricate

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 5 Typologies of hazardous waste that have been vitrified

Waste	Comments
BF, BOF, and other steelwork slag	Direct vitrification of BF [47]
	60 wt% BOF, additions of sand (glass formation) and Na_2O (workability) [48]
	Additions of bauxite residues, limestone, sand [50]
Zn hydrometallurgy slag	30–60 wt% of glass cullet or granite scraps and mud; high Fe content (>20 wt%) in the final glass [51, 52]
MSW fly ash	Addition of SiO_2 (10 wt%) to promote glass formation, MgO (5 wt%) to promote diopside formation upon conversion to glass-ceramic [61]
	Addition of 40 wt% SiO_2 and Na_2CO_3 powders, high Fe content leading to a nonhomogeneous glass [67]
	Addition of (30 wt%) waste from feldspar production [68]
	Ash rich in Pb; maximum ash content 60 wt%; addition of 40 wt% SiO_2 (glass former oxides in the ash <4 wt%) and Na_2CO_3 [67, 75]
Coal fly ash	Additions of Na_2CO_3 (to improve workability) and $CaCO_3$ (to improve chemical durability) [70]
Steel plant fly ash (+ MSW fly ash)	MSW fly ash gives chemical constituents of glass; addition of cullet (workability); and steel plant fly ash (crystallization) [73, 74, 76]
EAFD from stainless and carbon steel plants	55 wt% of glass cullet and sand; evaporation of Zn from the batch to increase the stability of the produced glass [69]
APC residues	Addition of silica (22 wt%) and alumina (8 wt%); soluble salt evaporated upon plasma melting [84]
Sewage sludge ash	Addition of limestone to obtain glass compositions prone to the formation of calcium alumino-silicates [87]
Mud from Bayer process (bauxite extraction)	Vitrification of a mixture of mud, coal fly ash, residues from the polishing of porcelain stoneware tiles and $CaCO_3$ [78]
Mud from Be extraction process	Additions of Na_2CO_3, Na_2SiF_6, and Na_3FeF_6 [90]
Dredging spoils	Addition of glass cullet (20 wt%) to promote glass formation [85, 86]
Asbestos-containing residues	Addition of K_2CO_3, $MgCO_3$, Ca-phosphate [91, 92]
Cullet	Composition corrections of batches comprising other types of inorganic waste (vast literature on the use of soda-lime-silica glass; limited experiences with Ba-Sr glass from the front part of CRTs) [96, 97]

high-value products, that is, by the introduction of a "valorization" step in the overall process. This opportunity depends on the fact that glass is a very versatile material and allows for the manufacturing of a wide range of products suitable for various applications. The use of waste-derived glass in a mass market application is highly attractive since it would enable a safe disposal of a large quantity of waste, even when the waste glass contributes only to a limited extent to the final composition of the material, with the additional benefit of reducing the use of natural raw materials. A major drawback of the approach is represented by the fact that glasses obtained from waste have often widely varying composition (and thus rheological behavior). Therefore, the process parameters of the selected production generally need to be reassessed for each starting

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 6 Typical composition of different waste (wt% to one decimal place)

Waste	SiO_2	Al_2O_3	Fe_xO_y	MgO	CaO	Na_2O	K_2O	ZnO	PbO	Other[a]	C+LOI
BF, BOF, and other steelwork slag [47, 50]	7.6–35.5	0.9–11.5	0.3–36	6.2–11.6	39–43.5	–	–	–	–	1.3–5	–
Zn hydrometallurgy slag [51]	3.7	0.3	49.3	0.2	0.1	–	–	5.6	3.6	–	37.2
MSW fly ash [10]	2.5–52.1	0.5–18	0.3–25.6	0.6–4.6	5.8–44.5	1.2–33	0.3–17.3	0.2–1.7	0–0.5	0.9–41	1.2–17.1
Coal fly ash [103]	15.2–66.2	11.4–29.6	3.8–43.5	0.6–9	2.6–23.7	0.2–5.1	0.3–2.1	<1.4	<0.1	<6.5	<41
EAFD from stainless and carbon steel plants [69]	4.4–5.9	0.7–1.5	24.2–52.8	5.2–9.6	7.5–20.7	0.9–6.6	1–1.7	7.6–13.8	<0.5	6.7–23.1	–
Sewage sludge ash [10]	14.4–57.7	4.6–26	2.7–24.6	0.9–4.2	4.1–38	0.1–1.2	0.1–2.7	0–0.5	–	0.6–26.7	0.3–15.1
Mud from Bayer process (bauxite extraction) [78]	7.8	17.1	44.1	0.6	11.7	3.2	0.1	–	–	5.6	9.8
Mud from Be extraction process [90]	75–80	16.9–19	0.3–9	0.1–0.2	0.1–0.2	0.3–2.6	–	–	–	1–1.7	–
Dredging spoils (calcined) [85]	42.4	12.9	5.2	8.2	25.8	1.5	2.3	–	–	0.9	–
Asbestos-containing residues [92]	30	3.9	2.5	8	33	0.3	0.3	–	–	0.3	21.5
CRT panel glass [93, 104]	58.9–60.7	3–1.7	0–0.1	0–0.9	0.1–1.7	7.5–8.1	6–6.9	0–0.1	0–3.4	18.4–18.5	–
CRT funnel glass [93, 104]	54.1–55.5	1.8–4.1	0–0.2	0–1.3	2.7–3.5	6.1–6.2	6.1–8.2	0–0.1	12–22	1.5–5.5	–

[a] B_2O_3, P_2O_5, MnO, BaO, SrO, ZrO_2, TiO_2, SO_3, Cl, F, etc.

material. Moreover, glasses from waste can be used only in applications where a high transparency is not required, because of the large amount of transition metal ions they contain, leading to dark-colored products.

Glass-Ceramics by Nucleation and Growth

Glass-ceramics represent a vast range of materials obtained by controlled crystallization of a glass of selected composition; the overall process leads to materials often possessing outstanding properties, such as high hardness and mechanical strength, a thermal expansion coefficient adjustable in a wide range of values (from negative to more than $12 \times 10^{-6}\ ^{\circ}C^{-1}$), high refractoriness, high chemical durability, and excellent dielectric properties [44]. The most valuable glass-ceramics, also known as "technical glass-ceramics," can be produced only from base glasses with a carefully controlled chemical composition, obtainable from particularly refined raw materials. However, the glass-ceramic technology has been applied to glasses obtained from waste since the early 1960s, that is, quite soon after the discovery of glass-ceramics occurred [44]. The significant variability in composition of the inorganic residues may be accommodated by using mixtures of different waste materials (changes in the waste ratio could compensate variations in the composition of a single waste) and by considering not particularly sophisticated applications, such as the manufacturing of tiles for the building industry [103].

Classical glass-ceramic technology relies on a double step treatment of a previously formed glass object (shaped into the desired form), corresponding to the nucleation of a crystal species within the base glass, favored by the separation of some glass components (such as Ag or Au colloids, or oxides like TiO_2 and ZrO_2), and to the crystal growth. These components are generally added to the formulation of the base glass, and are referred to as "nucleating agents." The base glass is heated first to the temperature of maximum nucleation and then to the temperature of maximum crystal growth (slightly higher than the previous one), with a holding time at each temperature, before cooling. These temperatures are different for each glass composition, and need to be determined precisely using, for instance, Differential Thermal Analysis (DTA).

The nucleation and growth heat treatment, often termed "ceramization," is the basis of a well-established production of glass-ceramics from waste glasses, known as "Slagsitalls," [44] and "Slagceram" [45]. Sheeted and pressed Slagsitalls have been produced for the last 40 years in more than 20 billion square meters and used in construction, chemical, mining, and other branches of industry. The base glasses for both Slagsitalls and Slagceram products belong to the systems $CaO–Al_2O_3–SiO_2$ and $CaO–MgO–Al_2O_3–SiO_2$, and are obtained from slags of ferrous and non-ferrous metallurgy, ashes and waste of mining and chemical industries, with minor compositional adjustments via addition of network-forming oxides (mainly SiO_2 and Al_2O_3) to form a stable glass upon cooling. Tiles are obtained mainly by a rolling process of a melt containing more than 50% of waste material, followed by crystallization. The addition of nucleating agents helps to achieve a uniform crystal growth under specific heat treatment conditions. The products obtained have very good mechanical strength and excellent abrasion resistance, due to the high percentage of crystals distributed uniformly in the whole volume and whose sizes range from 0.1 to 1 μm. Calcium silicate (wollastonite, $CaO \cdot SiO_2$) and calcium feldspar (anorthite, $CaO \cdot Al_2O_3 \cdot 2SiO_2$) are generally the main crystal phases, with other silicates and alumino-silicates (pyroxenes, i.e., rather complex chain silicates, expressed by the general formula $XY(Si,Al)_2O_6$, where $X = Na^+, Ca^{2+}, Fe^{2+}, Mg^{2+}$, etc., and $Y = Mg^{2+}, Fe^{2+}, Fe^{3+}, Al^{3+}, Cr^{3+}, Ti^{4+}$, etc., [106] or gehlenite $2CaO \cdot Al_2O_3 \cdot SiO_2$ and its solid solutions) present as secondary phases. However, depending on the composition, the secondary phases may in some cases replace the main ones and vice versa [44].

Microcrystalline products obtained from extensive nucleation (i.e., containing a very large number of nuclei) possess of very remarkable mechanical properties, even when produced from waste. Boccaccini et al. [58] showed an almost threefold increase of bending strength (from 90 to 240 MPa) and fracture toughness (from 0.6 to 1.7 MPa $m^{0.5}$) for a glass-ceramic with respect to the parent glass, produced from vitrification of MSW ash. Oveçoglu [75] produced slag-based glass-ceramics with a high bending strength (>300 MPa) and an excellent fracture toughness (5.2 MPa $m^{0.5}$). Peng et al. [107] demonstrated the feasibility of glass-ceramics with nano-sized

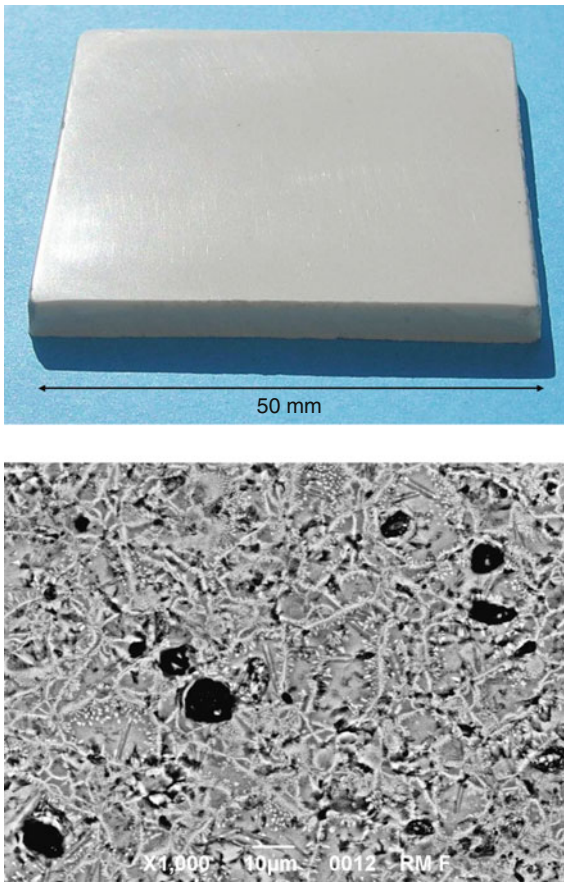
crystals (<200 nm), from the controlled crystallization of a glass from high alumina coal fly ash.

It must be observed that waste-derived glasses usually have an inherent strong tendency to devitrification, attributable to the presence of specific components in the formulation of waste, and therefore there is no need to add nucleating agents to the batch. The separation of magnetite (Fe_3O_4) crystals is particularly significant in iron-rich waste glasses. Karamanov and Pelino observed the dependence of crystallization on the ratio Fe^{3+}/Fe^{2+} [108, 109]. They showed that the crystallization of iron-rich glasses begins with the separation of small magnetite crystals, but the surface oxidation of Fe^{2+} to Fe^{3+} causes a change of the chemical composition, with the formation of hematite (Fe_2O_3), thus decreasing the total amount of crystal phase and changing the reaction order of the crystallization process. Bloomer et al. [110] also showed that oxidized glass from dc arc melting of highly iron containing waste were highly durable, since oxidative conditions promote the dissolution of heavy metals, but poorly prone to devitrification. Iron-rich waste glasses are known also for the opportunity of obtaining glass-ceramics with interesting functional properties: Romero-Perez et al. [63] showed that with Fe_2O_3 contents in the base glass superior to 26%, the precipitated magnetite particles were large enough to exhibit full magnetic order, so that glass-ceramics showed ferromagnetic behavior. Compositions with low waste introduction, up to 18% Fe_2O_3 concentration, behaved in paramagnetic way, due to iron dissolution in the glass matrix; intermediate compositions showed super-paramagnetic behavior due to fine precipitated magnetic clusters. Fe_2O_3 is also interesting for its interaction with sulfur: Susuki et al. [87] showed that, due to the presence of Fe_2O_3, sulfur and carbon, iron sulfide, FeS, could be formed and act as crystal nucleating agents for the precipitation of anorthite. Sulfides are also known to control the color of Slagsitalls: the addition of ZnO turns the color of glass-ceramics from grayish black, given by FeS or MnS, to white, due to the formation of ZnS (together with FeO or MnO) [44].

Glass-Ceramics by Sinter-Crystallization

The above described nucleation/crystal growth step may be difficult to control and economically expensive. A major drawback concerns the presence of defects in the glass articles, like pores, which remain in the glass-ceramic, causing a decrease in the mechanical properties. The evolution of gas bubbles from the glass melt requires high temperatures and long holding times, that is, a carefully controlled refining step. This refining is particularly complicated with waste glasses, which are usually dark and feature a low thermal conductivity by radiation, due to the amount of heavy metals. A further detrimental issue for glass-ceramics obtained by the traditional route is their visual appearance, which is generally rather inferior to that of natural stones and traditional ceramics. With a sintering approach, the problems of defects and visual appearance are generally avoided. In fact, when applying the sintering route, there is no need to refine the melt before casting into a frit, thus reducing cost and gaseous emissions. The ground glass powder is subsequently heated to a certain temperature, at which viscous flow sintering of glass powders occurs together with crystallization. The densification operated by the viscous flow greatly reduces the presence of pores; the crystallization, generally starting at the contact points between adjacent glass granules, [44] gives a pleasant marble-like visual appearance to the products (see Fig. 3a).

The simultaneous sintering and crystallization treatment is known as sinter-crystallization, [111] and it has been exploited commercially since the 1970s, for the manufacturing of the well-known Japanese, wollastonite-based, “Neoparies” tiles for the building industry [44]. The approach has been greatly reevaluated in the last 20 years for the conversion of waste glasses into useful products, again in the form of tiles, mainly due to the work of Gutzow and Karamanov [49]. In fact, remarkable advantages in both the vitrification and ceramization steps are achievable. As mentioned above, refining is not needed, so that the vitrification may be conducted in small plants and in particularly short times, favoring the immobilization of components, which could vaporize with longer heat treatments. Furthermore, a relatively high degree of crystallization may be achieved in very short times; the surface of glass is in fact a preferred site for crystallization, [112–115] and thus ground glass is easier to devitrify than bulk glass with the same composition, and nucleating agents are not needed. In some cases, the holding time at the sintering temperature may not exceed 30 min, being also accompanied by very fast heating rates (even “direct heating” is possible, that is



Vitrification of Waste and Reuse of Waste-Derived Glass. Figure 3

(**a**) Visual appearance of a sintered sanidine-based glass-ceramic tile; (**b**) typical microstructure of a sintered glass-ceramic from vitrified waste, containing red mud from the Bayer process (the lighter crystals correspond to pyroxenes)

the direct insertion of glass powder compacts in the furnace directly at the sintering temperature), thus configuring a "fast sinter-crystallization" process [61, 116]. The base glasses for the manufacturing of sintered glass-ceramics have similar chemical compositions, except for lack of nucleating agents, to those of glass-ceramics from waste glasses obtained by conventional nucleation and growth. Pyroxenes, wollastonite, and anorthite (with solid solutions) are very common crystal phases (see Fig. 3b). However, the remarkable nucleation activity of fine glass powders (<40 μm) has been found to enable the very unusual precipitation of alkali feldspars and feldspathoids, such as sanidine and nepheline, as main crystal phases [96, 97].

The sinter-crystallization process relies on a quite complicated balance between viscous flow sintering, surface crystallization, and even bulk crystallization, that is, crystallization operated by the separation of components acting as nucleating agents. This balance is sensible to many conditions, for example, the oxidation state and the heating rate. Starting from an iron-rich waste glass, Karamanov et al. [117] observed that the addition of C (1.5–2%) to the glass batch increased the magnetite phase and enhanced the crystallization rate. Bernardo et al. [78] starting from a base waste glass with a low Fe^{2+}/Fe^{3+} ratio, observed that magnetite was promoted by oxidation, more sensible for fine glass powders (<40 μm) than for coarse ones (<80 μm). Karamanov et al. [118, 119] reported that the balance between surface crystallization and bulk crystallization is strongly affected by the heating rate: low heating rates favor bulk crystallization, and sintering may be inhibited by the crystal phase, causing incomplete densification. High heating rates favor sintering so that low porosity remains in the material; however, the amount of crystal phase formation is lower, because crystallization occurs only at the surface. It has been shown in many papers [78, 86, 96, 97, 116] that, in the presence of fine glass powders (<40 μm), the crystallization may be achieved right at the temperature of the crystallization exothermic peak in the DTA plot of the same powders. More recent investigations, [61] however, highlighted that optimum crystallization is achievable only if the crystallization peak is located at a temperature suitably higher than that corresponding to the dilatometric softening point, that is, the temperature at which viscous flow becomes appreciable [120]. If the temperature difference is limited, the obtained glass-ceramics are remarkably porous and improvements in the densification are achievable only by increasing the sintering temperature and the heating rate (direct heating, as described above, enables sintering of the powders before the crystallization can "freeze" the viscous flow, due to the very large increase in viscosity associated to crystal precipitation).

Glass-Ceramics by the Petrurgic Method

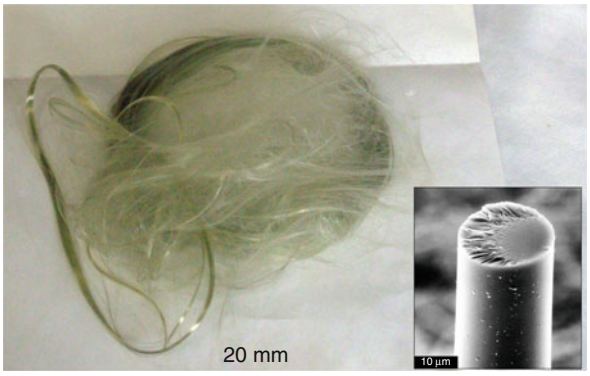
Glass-ceramics from waste may be produced by a third method, known as the "petrurgic method" [121], named in this way because of the similarity with the process of crystallization of natural rocks. This method has actually been applied since the 1970s, with the development of "Silceram" ceramics from metallurgical slags [103]. In this process, crystals nucleate and grow directly upon cooling of glass from the melting temperature, with an intermediate temperature holding stage, [66] which can sometimes be avoided. Francis et al. [83] reported the feasibility of crystallization upon controlled cooling (from 1°C to 10°C/min) of glasses obtained from mixtures of coal ash and soda lime glass melted at 1,500°C, without any intermediate step. The cooling rate is a dominant factor in controlling the formation and morphology of the crystal phase, particularly in relation to iron containing raw materials. Faster cooling rates allow for the formation of magnetite, with samples exhibiting magnetic properties, while slow cooling rates cause the formation of plagioclase and augite.

"Glass-Ceramics" by Direct Sintering of Silicate Waste

The above mentioned techniques for the manufacturing of glass-ceramic all involve the previous formation of a glass. However, some interesting sintered materials may be produced by mixing glass cullet, of various origins, with silicate waste, leading to components which cannot, strictly speaking, be termed glass-ceramics, since they are not produced by the (controlled) crystallization of any parent glass. Francis et al. [122] as an example, produced glass-ceramics from high iron containing coal ash, adding Pyrex glass powder. The presence of iron led to soft magnetic materials, due to ferrite phases formation depending on the ash/glass ratio. A similar approach was followed by Fidancevska et al. [123]. In these cases, the crystallization is associated to the interaction between recycled glass and the silicate waste.

Fibers

Whereas glass-ceramics have constituted so far the main application for waste-derived glasses, glass fibers for composites and thermal insulation represent also a valid opportunity (see Fig. 4). Scarinci et al. [124] obtained glass fibers from glasses derived by the vitrification of MSW ashes and sludge excavated from the lagoon of Venice, with glass cullet added as melting aid. The glass fibers displayed a good tensile strength (maximum value of 1.6 GPa) and an elastic modulus up to 75 GPa. Glass fibers for reinforcement of plastics or bituminous materials, possessing physical and chemical properties comparable to those of commercially available materials, were also produced from a variety of waste [125]. The production of glass fibers is also a promising application for panel glass from cathode ray tubes (CRT), as shown by Hreglich et al. [126]. The addition of about 15 wt% of panel glass to the typical batch composition for type A glass fibers does not significantly modify the working range temperature, enabling the production of fibers of comparable quality to commercially available ones. These fibers could find application in textiles for radiation protection. Marabini et al. obtained glass wool by melting talc and cromite mine tailings (with basalt rock employed as an additive) at 1,350°C and running the fluid onto a rapid rotating heated plate; the obtained glass fibers were subsequently crystallized [127].



Vitrification of Waste and Reuse of Waste-Derived Glass. Figure 4

Glass fiber tow derived from waste (Courtesy of Dr. Roberto Falcone, Stazione Sperimentale del Vetro, Venice, Italy – http://www.spevetro.it). Inset: a typical SEM micrograph of a single fiber

Cellular Glasses and Glass-Ceramics

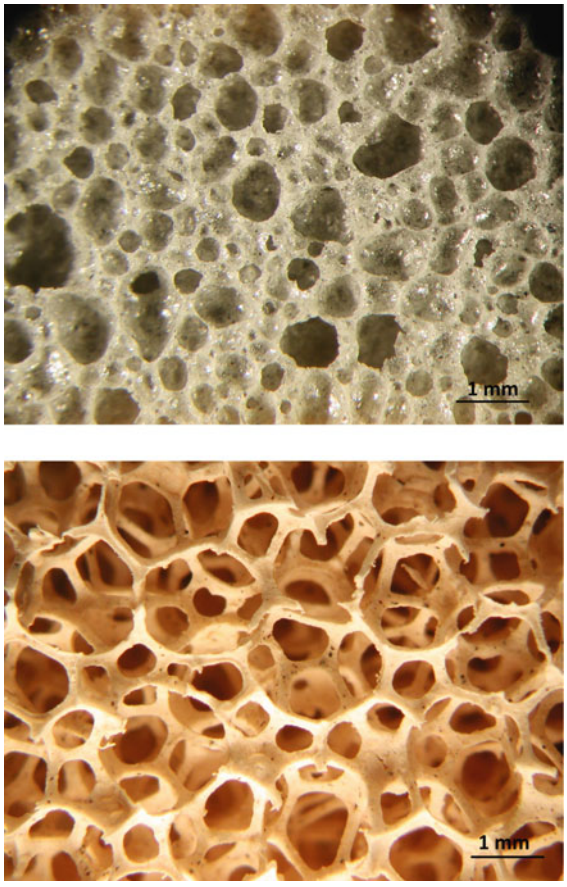
Glass foams represent a particularly interesting type of components for thermal and acoustic applications [128].

They are produced in limited quantities due to high processing costs, but their characteristic of un-flammability, thermal and chemical stability will lead to an increased use in substitution for organic foams. Moreover, the closed cell structure makes the material watertight and an efficient barrier against soil humidity. The production of glass foams may follow two distinct processes: [128] the first, dating back to the 1930s, consists of the direct introduction of gases ("blowing") into molten glass; the second one, much less expensive, is based on the viscous flow sintering of fine glass powders, which creates a pyroplastic mass, which is foamed by the action of specific powder additives (foaming agents), owing to decomposition or oxidation reactions. The decomposition reactions involve carbonates and sulfates, while oxidation reactions are due to the interaction of carbon-containing species (C, SiC) with oxygen, coming mainly from the atmosphere of the sintering furnace. The adoption of a sintering approach paved the way for the use of glass not specifically designed for foam production; significantly, the sintering approach led to the extensive use of cullet for this application. Soda-lime glass is the common raw material; however, a number of recent investigations showed that it is possible to fabricate foams using other glasses, such as CRT glass [93]. The low characteristic temperature of these glasses enables foaming at particularly low temperature (even below 750°C), through the decomposition of added calcium carbonate. Furthermore, this foaming procedure has the significant advantage of producing chemically stable foams, avoiding any risk of reduction of heavy metal oxides into metal colloids, which is often experienced when using reducing compounds (SiC or TiN) [129].

Glass foams represent an interesting way to directly use industrial residues, without converting them into a glass, with the addition of recycled soda-lime glass. Brusatin et al. [130] used fuel oil-derived fly ashes (with SiC as foaming agent), processing at temperatures between 800°C and 900°C, producing foams with a crushing strength of about 3.5 MPa and a density of ~0.35 g/cm^3. Bernardo et al. [94] used a SiC-containing waste, deriving from the polishing operations of artistic glass articles, as the only foaming agent (although the oxidation of SiC was favored by the addition of MnO_2), obtaining very high porosities (up to 92%); the heavy metal oxides present in the waste (mainly lead and cadmium deriving from artistic glass) were found to be immobilized in the glass matrix. Fernandes et al. also employed waste as raw materials containing a foaming agent, using ash produced after the burning out of SiC-based abrasive paper [131].

The foaming of waste-derived glasses is quite complicated, due the above discussed tendency of these glasses to crystallize upon heating. Significantly, a porous structure is associated to a high specific surface, which enhances surface nucleation. If extensive crystallization occurs during foaming, problems in the homogeneity and reproducibility of the overall foam morphology arise. This issue may be overcome by using a combined approach, that is, by the foaming of mixtures of soda-lime glass and a glass undergoing crystallization, [132] or again of soda-lime glass and selected waste (like in the case of "glass-ceramics" from direct sintering of silicate waste) [133]. In this case the crystallization may actually be useful, since it enhances the mechanical properties. A similar situation is found when using mixtures of soda-lime glasses with cullet of other origin, more difficultly recycled but quite prone to crystallization (e.g., glass residue from the manufacturing of glass fibers, having a $CaO–Al_2O_3–SiO_2$ composition similar to that of many waste glasses used for the preparation of glass-ceramics), as observed by Bernardo et al. [95] (see Fig. 5a).

The remarkable crystallization of waste glasses may be even advantageous when other types of cellular materials are desired. Common glass foams are mostly closed cell foams, thus maximizing the thermal insulation; on the other hand, this makes them unsuitable to filtering applications. Open-celled glass-ceramics may be obtained by the mixing of glass powders with polymeric microspheres or by the deposition of slurries onto polyurethane sponges, [134] followed by the burn-out of the sacrificial polymers and sinter-crystallization (see Fig. 5b). The crystallization, enhanced by the high amount of free glass surface, "freezes" the structure modeled by the polymeric templates, greatly limiting the viscous collapse. A similar concept was reported by Fidancevska et al., who used slurries composed of mixtures of coal fly ash and recycled glasses to coat a polyurethane foam or to envelop sacrificial carbon fibers [123].



Vitrification of Waste and Reuse of Waste-Derived Glass. Figure 5
(**a**) Glass foam obtained from a mixture of recycled glasses; (**b**) cellular glass-ceramic produced by sinter-crystallization of fine glass powders deposited on a polyurethane sponge

Glass and Glass-Ceramic Matrix Composites

The development of glass matrix composites from waste was pioneered by Boccaccini et al. [135] who reported the introduction of up to 20 vol% low-cost alumina platelets in a glass matrix developed by sintering borosilicate glass cullet mixed with fly ash, leading to an increase in Young's modulus, modulus of rupture, hardness, and fracture toughness with platelet volume fraction. The reinforcement is mainly intended to improve the fracture toughness of the matrix, the most important limit for extensive structural use applications of glass-based products. Ferraris et al. [136] following the same approach, reported the introduction of solid waste from aluminum foundry.

Glass-ceramic composites can also be easily obtained using the sinter-crystallization approach, by adding the reinforcement directly to glass powders, subsequently heated. Bernardo et al. [96] prepared composites with a bending strength of 163 ± 14 MPa and a fracture toughness of 1.9 ± 0.1 MPa·$m^{0.5}$, by the addition of up to 15 vol% alumina platelets to a waste glass capable of sinter-crystallization and leading to a nepheline-based glass-ceramic matrix. Appendino et al. [137] and Aloisi et al. [138] found similar results with a glass from MSW incinerator fly ash mixed with alumina waste. However, the addition of alumina platelets is more significant because it gives origin to an alternative material to glass-ceramics, rather than being used for the reinforcement of glass-ceramics. Pressureless sintered alumina-reinforced glass matrix composites have in fact been prepared from CRT glasses [139] and from waste glass not prone to sinter-crystallization [140]. In both cases, the overall mechanical properties (e.g., a bending strength of about 100 MPa) are very close to those exhibited by Slagsitalls, although processed at much lower temperature (e.g., only 650°C for the composites based on CRT glasses). The alumina reinforcement may also be exploited even for "glass ceramics" from direct sintering of silicate waste; Rozenstrauha et al. [141] reported in fact a significant enhancement in fracture strength (from 57 to 97 MPa) with a 20 wt% addition of platelets.

Further Applications

A number of secondary applications for waste-derived glass, in addition to those previously presented, is available. Minor applications include products such as tableware, optical glass, insulators, and solid fertilizers. Hreglich et al. [126] successfully produced colored tableware glass using panel glass from cathode ray tubes, with a limited addition of sodium nitrate as a fining material. Diaz et al. [55] prepared transparent optical glass from the vitrification of geothermic waste, added with Pb_3O_4 and K_2CO_3, acting as modifiers and stabilizers. Saccani et al. [142] showed that a glass derived from the melting of MSW incinerator ashes and cullet possesses an electrical resistivity comparable

to that of E glass, so that it could be employed for high voltage insulators. Barba et al. [143] produced low cost phosphorous and potassium containing glass-ceramic frits by melting phosphates and feldspar with bones and glass cullet and pouring the melt into water. The leaching behavior of such frits suggests their application as solid fertilizers. A very similar approach was recently followed by Dall'Igna et al. [92] using vitrified asbestos-containing waste as a low cost raw material.

Glazes, tiles, road pavement, and concrete represent an important opportunity for using waste-derived glasses, because of the very large volume of production of such articles. Waste-derived glass may be used as inert aggregate in concrete, [144] asphalt concrete for road pavements, [145, 146] filler in plastics and paints etc. Finally, some studies illustrated the possibility of producing high-performance abrasive powders (Vickers' hardness >9 GPa) from the crystallization of vitrified electric arc furnace dust and foundry sand [147].

The use of waste-derived glass in the ceramic industry needs a more detailed discussion. Most of the applications of waste glasses presented in the previous paragraphs are related to products made almost completely of glass from waste, but they are of interest only to a limited market. The ceramic products discussed below, on the contrary, do not feature glass as their main constituent, but have a mass market. The manufacturing of ceramic tiles is particularly interesting due to their higher added value. Waste-derived glasses may be used in both coatings and body formulation, acting as a sintering aid. Glazes were developed from panel glass cullet [126], metallurgical slags [148], or granite waste [149]; vitrified MSW ash were employed as a sintering aid in the fabrication of porcelain stoneware or other traditional ceramics [86, 150] in analogy to glass cullet, for which a vast literature is already available [98–102]. It was observed that vitrified MSW ash in the fabrication of porcelain stoneware may cause an undesirable darker coloration with respect to the base body, with a slight coarsening in the tiles planarity, and this limited the maximum amount of waste-derived glass that can be used. Recent investigations [151] showed that this conclusion is probably premature. Firstly, the unpleasant coloration may be modified by the use of pigments; secondly, waste-derived glasses may represent an opportunity for the development of a new type of porcelain stoneware, fired at much lower temperatures than the traditional ones and with a much reduced use of natural raw materials. For instance, feldspar sands were completely replaced by a $CaO–Al_2O_3–SiO_2$ waste-derived glass, [151] obtaining a ceramic (fired at 1,000°C) with the same mechanical properties than that produced from traditional raw materials (fired above 1,150°C). The key advantages in the replacement of feldspars are the fact that the waste glass may provide a liquid phase by softening at much lower temperature than that required by the melting of feldspars, and the fact that some crystallization occurs, leading to an increase of viscosity that stabilizes the shrinkage. The complete replacement of feldspar with recycled soda-lime glass and CRT glasses has been recently proposed [104]. The stabilization of shrinkage could be caused by a viscosity increase, operated by the addition of calcium hydroxide (for soda-lime glass) promoting the formation of wollastonite, or by the addition of alumina platelets (for CRT glasses). In both cases, the firing temperatures were particularly low (880–920°C and 750–775°C, respectively).

In Table 7 are reported the main typologies of products that have been so far fabricated using glass produced from waste. The references quoted refer to the selected representative papers, for example, those reporting values for some engineering properties of the material (mechanical strength, hardness, etc.) or significant advances in the processing.

Future Directions

The vitrification of waste is a technology that is successful and suitable for the inertization of a large variety of waste, differing in composition and form; however, it still has to be proven economically viable on any substantial scale due to the substantial energy cost required to produce the high temperature necessary for obtaining a workable and homogeneous melt. If the primary economic driver toward vitrification is only the avoidance of landfill cost, then this additional energy-intensive process is not justifiable per se. However, when in some instances the hazardous nature of the waste is such that immobilization is absolutely required to meet regulation requirements or, as in the case of contaminated land remediation where the cost of vitrification can be sometimes comparable to alternative cleanup costs, sufficient incentives to pursue

Vitrification of Waste and Reuse of Waste-Derived Glass. Table 7 Main products fabricated from waste-derived glasses

Product fabricated	Main typologies of waste used	Comments	Selected references
Glass-ceramic tiles (by nucleation and growth)	MSW incinerator ashes; metallurgical slag; coal fly ash; cullet	Nucleation triggered by additives to the glass batch (e.g., TiO_2 or P_2O_5) or by separation of iron oxide already present in the waste. Most treatments require a two-step (nucleation and growth) heat treatment for crystal phase development	[58, 63, 75, 87, 107]
Glass-ceramic tiles (by sinter-crystallization)	MSW incinerator ashes; dredging spoils; mixtures of various waste	No need for second heat treatment; influence of oxidation state and heating rate; possibility of "fast sinter-crystallization"	[49, 61, 68, 116–119]
Glass-ceramic tiles (by petrurgic method)	Coal fly ash; cullet	Crystalline material forms upon cooling, with or without intermediate temperature holding stage	[83]
Fibers	MSW incinerator ashes; dredging spoils; cullet	Suitable for reuse of CRT cullet	[124, 125, 127]
Porous glasses and glass-ceramics	Cullet	Viscosity of glass affects choice of foaming agents; crystallization to be limited (in closed cell foams) or exploited (in open-celled foams)	[93, 94, 131, 134]
Glass and glass-ceramic matrix composites (tiles)	MSW incinerator ashes; Al foundry waste; coal fly ash; cullet	Alumina platelets as main reinforcement. Suitable for reuse of CRT cullet	[96, 137–139]
Frits (for glazes)	Metallurgical slags; granite waste	Difficulties in controlling color. Suitable for crystalline coatings	[148, 149]
Porcelain stoneware tiles	MSW incinerator ashes; cullet; mixtures of various waste	Possible complete replacement of feldspars and reduction of firing temperature; controlled shrinkage (by crystallization)	[104, 150, 151]
Solid fertilizers	Bones; cullet; asbestos waste	Very specific compositions need to be achieved	[92, 143]
Aggregate in concrete (cementitious or bituminous)	Cullet	Low cost and low bulk density; decorative; problems with debonding of glass particles	[144–146]

this route exist. Moreover, the reuse of the glass or glass ceramic produced from the vitrification of waste, either as new products or as raw materials for other processes, would contribute to further offset the overall cost of the technology.

The forecasted future developments concern a rationalization and specialization of the melting technologies, based from one hand on a more exact knowledge of the initial batch (composition, physical state, particle size, toxicity, etc.), and the other hand on precise selection rules (based on economic or plant criterions) of the off-gas purification systems. Alternatively, the development of a multipurpose melter (such as the cold-crucible induction melter, recently devised for the treatment of nuclear waste) could meet the growing demand of successfully vitrifying waste with vastly different chemical-physical characteristics, however increasing the cost and requirements relative to the gas depuration system. Changes in legislatory regulations would also prove crucial toward aiding to establish vitrification as not only the safest, but also the most viable hazardous waste management process.

The present trend, which will be more and more continued in the future due to economic reasons, is toward the fabrication of marketable products using glass obtained from waste, such as glass-ceramics, tiles

for flooring or lining, foam glass, insulating or reinforcing fibers, glass or glass-ceramic composites. The scientific and technological challenge will be to produce components possessing the chemical, mechanical, and esthetic characteristics required by the specific application.

Bibliography

Primary Literature

1. Roth G, Weisenburger S (2000) Vitrification of high level liquid waste: glass chemistry, process chemistry and process technology. Nucl Eng Des 202:197–207
2. Park J-K, Song M-J (1998) Feasibility study on vitrification of low-and intermediate-level radioactive waste from pressurized water reactors. Waste Manag 18:157–167
3. Sakai S, Hiraoka M (2000) Municipal solid waste incinerator residue recycling by thermal processes. Waste Manag 20:249–258
4. US Environmental Protection Agency (1994) 1994 EPA contaminated sediment management strategy, EPA 823-R-94-001. Office of Water, US Environmental Protection Agency, Washington
5. SITE (Superfund Innovative Technology Evaluation) Emerging Technology Bulletin (1995) Ferro Corporation waste vitrification through electric melting, U.S. EPA/540/F-95/503. US Environmental Protection Agency, Cincinnati
6. Nechvatal TM, Jansen TJ (1996) Converting paper mill sludge or the like. US Patent 5,549,059. Assignee: Minergy
7. Buelt JL, Oma KH, Eschbach EA (1994) Apparatus for in situ heating and vitrification. US Patent 5,316,411, 31 May 1994
8. U.S. Environmental Protection Agency (1995) Geosafe Corporation in situ vitrification. innovative technology evaluation report. Risk reduction engineering laboratory, Office of Research and Development. Report EPA/540/R-94/520
9. Poiroux R, Rollin M (1996) High temperature treatment of waste: from laboratories to the industrial stage. Pure Appl Chem 68:1035–1040
10. Bingham PA, Hand RJ (2006) Vitrification of toxic waste: a brief review. Adv Appl Ceram 105:21–31
11. http://www.epa.gov/epawaste/hazard/tsd/td/combustion.htm. Accessed 10 Feb 2009
12. Marra JC, Jantzen CM (2004) Glass: an environmental protector. Am Ceram Soc Bull 83(11):12–16
13. Baehr W (1989) Industrial vitrification processes for high-level liquid waste solutions. IAEA Bulletin 31(4):47–51
14. Buelt JL, Chapman C (1978) Liquid fed ceramic melter. Doc. N° PNL-2735, UC 70, U.S. Department of Energy
15. Jantzen C, Bickford DF, Brown KG, Cozzi AD et al (2000) Savannah river site waste vitrification projects initiated throughout the United States: disposal and recycle options. US Department of Energy, Office of Scientific and Technical Information, Oak Ridge
16. Roth G (1995) Atomwirtschaft 40(Jg3):174–177
17. Jouan A (2001) La vitrification des déchets, une contribution au respect de notre terre. Verre 7:20–27
18. US Environmental Protection Agency (1992) Handbook on vitrification technologies for treatment of hazardous and radioactive waste, report EPA/625/R-92/002. Office of Research and Development, Washington
19. Buelt JL (1997) Molten glass processes. In: Freeman HM (ed) Standard handbook of hazardous waste treatment and disposal, 2nd edn. McGraw-Hill, New York, pp 45–77
20. Wakamura Y, Nakazato K (1994) Recent trend of ash management from MSW incineration facilities in Japan. In: National Waste Processing Conference Proceedings ASME 91–96
21. Richards RS, Plodinec MJ (1998) Overview of current and emerging waste vitrification technologies. In: Proceedings of the XVIII international congress on glass, San Francisco, 5–8 July 1998. Paper No. A7-I (CD ROM). The American Ceramic Society, Westerville
22. Hollander H (1995) Vitrification of combustion ash residue for beneficial use. Solid Waste Technol 9:31–40
23. Terasawa Y, Yasuda S, Horizoe H, Sato J, Gotou Y (2001) Commercialization of MSW incineration system with direct ash melting by thermal cracking for high efficient generation of electricity, Mitsubishi Heavy Industries, Ltd. Tech Rev 38(2):82–86
24. Miyata H, Sadatsuka T (2005) Technology applicable to "Heat recovery facilities", Sanki Engineering. J Solid Liq Waste 35(9):43–44, in Japanese
25. Chapman C (1995) Earth melter. US Patent 5,443,618. Assignee Battelle Memorial Institute, Richland
26. Chapman C (1993) State-of-the-art of waste glass melters. In: Varshneya AK, Bickford DF, Bihuniak PP (eds) Ceramic Transactions v.29. American Ceramic Society, Westerville, pp 485–493
27. Park JK, Moon YP, Park BC, Song MJ, Ko KS, Cho JM (2001) Noncombustible waste vitrification with plasma torch melter. J Environ Sci Health A Tox/Hazard Subst Environ Eng 36:861–871
28. Tendler M, Retberg P, Van Oost G (2005) Plasma based waste treatment and energy production. Plasma Phys Controlled Fusion 47:A219–A230
29. Moustakas K, Fatta D, Malamis S, Haralambous K, Lozidou M (2005) Demonstration plasma gasification/vitrification system for effective hazardous waste treatment. J Hazard Mater 123:120–126
30. Park HS, Kim SJ (2005) Analysis of a plasma melting system for incinerated ash. J Ind Eng Chem 11:657–665
31. Kushnikov VV et al. (1995) Using an induction melter with a cold crucible for the immobilization of Plutonium. In: Plutonium Stabilization and Immobilization Workshop Proceedings, Washington, pp 319–326
32. Jouan A, Boen R, Merlin S, Pujadas V (1997) New development for medium and low level waste vitrification. In: Nuthos-5, Beijing, 14–18, April 1997
33. Ojovan MI, Lee WE (2003) Self sustaining vitrification for immobilisation of radioactive and toxic waste. Glass Technol 44:218–224
34. Karlina OK, Varlakova GA, Ojovan MI, Tivanski VM, Klimov VL, Pavlova GY, Dmitriev SA (2001) Ash and soil conditioning using exothermic metallic compositions. Mater Res Soc Symp Proc 663:65–70

35. Blackman WC (1993) Basic hazardous waste management. Lewis, Boca Raton
36. European Council (2000) European Waste Catalogue, Council Decision 2000/532/EC, Official Journal of the European Communities L226:3–24
37. Räbiger K, Keldenich K, Scheffer J (1995) Experience in operation of a pilot plant melting residual substances. Glastech Ber Glass Sci Technol 68:84–90
38. Frugifer P, Godon N, Vernaz E, Larché F (2002) Influence of composition variations on the initial alteration rate of vitrified domestic waste incineration fly-ash. Waste Manag 22:137–142
39. Piepel G, Redgate T (1997) Mixture techniques for reducing the number of components applied for modeling waste glass sodium release. J Am Ceram Soc 80:3038–3044
40. Besmann TM, Spear KE (2002) Thermochemical modeling of oxide glasses. J Am Ceram Soc 85:2887–2894
41. Kim C-W, Choi K, Park J-K, Shin S-W, Song M-J (2001) Enthalpies of chromium oxide solution in soda lime borosilicate glass systems. J Am Ceram Soc 84:2987–2990
42. Lapa N, Santos Oliveira JF, Camacho SL, Circeo LJ (2002) An ecotoxic risk assessment of residue materials produced by the plasma pyrolysis/vitrification (PP/V) process. Waste Manag 22:335–342
43. Colombo P, Brusatin G, Bernardo E, Scarinci G (2003) Inertization and reuse of waste materials by vitrification and fabrication of glass-based products. Curr Opin Solid State Mater Sci 7:225–239
44. Höland W, Beall G (2002) Glass-ceramic technology. American Ceramic Society, Westerville
45. Davies MW, Kerrison B, Gross WE, Robson MJ, Witchall DF (1973) Slag ceramics: a glass ceramic from blast-furnace slag. J Iron Steel Inst 208:348–370
46. Nakamura S (1976) Crystallized glass article having a surface pattern. US patent 3,955,989. 11 May 1976
47. Fredericci C, Zanotto ED, Ziemath EC (2000) Crystallization mechanism and properties of a blast furnace slag glass. J Noncryst Solids 273:64–75
48. Ferreira EB, Zanotto ED, Scudeller LAM (2002) Glass and glass-ceramic from basic oxygen furnace (BOF) slag. Glass Sci Technol 75:75–86
49. Karamanov A, Gutzow I, Chomakov I, Christov J, Kostov L (1994) Synthesis of wall-covering glass-ceramics from waste raw materials. Glastech Ber Glass Sci Technol 67:227–230
50. Gomes V, De Borba CDG, Riella HG (2002) Production and characterization of glass ceramics from steelwork slag. J Mater Sci 37:2581–2585
51. Pelino M (2000) Recycling of zinc-hydrometallurgy waste in glass and glass ceramic materials. Waste Manag 20: 561–568
52. Karamanov A, Taglieri G, Pelino M (1999) Iron-rich sintered glass-ceramics from industrial waste. J Am Ceram Soc 82(11):3012–3016
53. Pisciella P, Crisucci S, Karamanov A, Pelino M (2001) Chemical durability of glasses obtained by vitrification of industrial waste. Waste Manag 21:1–9
54. Diaz C, Valle-Fuentes FJ, Zayas ME, Avalos-Borja M (1999) Cordierite glass-ceramic from geothermic waste. Am Ceram Soc Bull 78:62–64
55. Diaz C, Gracia H, MaE Z, Espinoza FJ, Valle-Fuentes FJ (2000) Producing optical glass with geothermal waste. Am Ceram Soc Bull 79:57–59
56. Ferreira C, Ribeiro A, Ottosen L (2003) Possible applications for municipal solid waste fly ash. J Hazard Mater B96:201–216
57. Romero M, Rawlings RD, Rincón JM (1999) Development of a new glass-ceramic by means of controlled vitrification and crystallization of inorganic waste from urban incineration. J Eur Ceram Soc 19:2049–2058
58. Boccaccini AR, Kopf M, Stumpfe W (1995) Glass-ceramics from filter dusts from waste incinerators. Ceram Int 21:231–235
59. Cheng TW, Chen YS (2003) On formation of CaO-Al_2O_3-SiO_2 glass-ceramics by vitrification of incinerator fly ash. Chemosphere 51:817–824
60. Park YJ, Heo J (2002) Conversion to glass-ceramics from glasses made by MSW incinerator fly ash for recycling. Ceram Int 28:689–694
61. Bernardo E, Scarinci G, Edme E, Michon U, Planty N (2009) Fast-sintered gehlenite glass–ceramics from plasma-vitrified municipal solid waste incinerator fly ashes. J Am Ceram Soc 92:528–530
62. Romero M, Rawlings RD, Rincón JM (2000) Crystal nucleation and growth in glasses from inorganic waste from urban incineration. J Noncryst Solids 271:108–118
63. Romero M, Rincon JMa, Rawlings RD, Boccaccini AR (2001) Use of vitrified urban incinerator waste as raw material for production of sintered glass-ceramics. Mater Res Bull 36:383–395
64. Park YJ, Heo J (2002) Vitrification of fly ash from municipal solid waste incinerator. J Hazard Mater B91:83–93
65. Siwadamrongpong S, Koide M, Matusita K (2004) Prediction of chloride solubility in CaO–Al_2O_3–SiO_2 glass systems. J Noncryst Solids 347:114–120
66. Kim JM, Kim HS (2004) Glass-ceramic produced from a municipal waste incinerator fly ash with high Cl content. J Eur Ceram Soc 24:2373–2382
67. Kavouras P, Komninou Ph, Chrissafis K, Kaimakamis G, Kokkou S (2003) Microstructural changes of processed vitrified solid waste products. J Eur Ceram Soc 23:1305–1311
68. Karamanov A, Pelino M, Hreglich S (2003) Sintered glass-ceramics from municipal solid waste-incinerator fly ashes-part I: the influence of the heating rate on the sinter-crystallization. J Eur Ceram Soc 23:827–832
69. Pelino M, Karamanov A, Pisciella P, Crisucci S, Zonetti D (2002) Vitrification of electric arc fornace dusts. Waste Manag 22:945–949
70. Leroy C, Ferro MC, Monteiro RCC, Fernandes MHV (2001) Production of glass-ceramics from coal ashes. J Eur Ceram Soc 21:195–202
71. Kavouras P, Kaimakamis G, Ioannidis ThA, Kehagias Th, Komninou Ph, Kokkou S, Pavlidou E, Antonopoulos I, Sofoniou M, Zouboulis A, Hadjiantoniou CP, Nouet G, Prakouras A, Karakostas Th (2003) Vitrification of lead-rich solid ashes

from incineration of hazardous industrial waste. Waste Manag 23:361–371
72. Cheng TW (2003) Combined glassification of EAF dust and incinerator fly ash. Chemosphere 50:47–51
73. Barbieri L, Ferrari AM, Lancellotti I, Leonelli C (2000) Crystallization of (Na_2O-MgO)-CaO-Al_2O_3-SiO_2 glassy systems formulated from waste products. J Am Ceram Soc 83:2515–2520
74. Barbieri L, Corradi A, Lancellotti I (2000) Alkaline and alcaline-earth silicate glasses and glass-ceramics from municipal and industrial waste. J Eur Ceram Soc 20:2477–2483
75. Öveçoğlu ML (1998) Microstructural characterization and physical properties of a slag-based glass-ceramic crystallized at 950 and 1100°C. J Eur Ceram Soc 18:161–168
76. Barbieri L, Corradi A, Lancellotti I (2002) Thermal and chemical behavior of different glasses containing steel fly ash and their transformation into glass-ceramics. J Eur Ceram Soc 22:1759–1765
77. Barbieri L, Lancellotti I, Manfredini T, Queralt I, Rincon JM, Romero M (1999) Design, obtainment and properties of glasses and glass-ceramics from coal fly ash. Fuel 78:271–276
78. Bernardo E, Esposito L, Rambaldi E, Tucci A, Pontikes Y, Angelopoulos GN (2009) Sintered esseneite–wollastonite–plagioclase glass–ceramics from vitrified waste. J Eur Ceram Soc 29:2921–2927
79. Boccaccini A, Rawlings R (2002) Waste not – Producing glass-ceramics from waste materials. Mater World 10:16–18
80. Rincon JM, Romero M, Boccaccini AR (1999) Microstructural characterisation of a glass and a glass-ceramic obtained from municipal incinerator fly ash. J Mater Sci 34:4413–4423
81. Boccaccini AR, Petitmermet M, Wintermantel E (1997) Glass-ceramics from municipal incinerator fly ash. Am Ceram Soc Bull 76:75–78
82. Erol M, Demirler U, Küçükbayrak S, Ersoy-Meriçboyu A, Öveçoğlu ML (2003) Characterization investigations of glass-ceramics developed from Seyitömer thermal power plant fly ash. J Eur Ceram Soc 23:757–763
83. Francis AA, Rawlings RD, Boccaccini AR (2002) Glass-ceramics from mixtures of coal ash and soda lime glass by the petrurgic method. J Mater Sci Lett 21:975–980
84. Amutha Rani D, Gomez E, Boccaccini AR, Hao L, Deegan D, Cheeseman CR (2008) Plasma treatment of air pollution control residues. Waste Manag 28:1254–1262
85. Bernstein AG, Bonsembiante E, Brusatin G, Calzolari G, Colombo P, Dall'Igna R, Hreglich S, Scarinci G (2002) Inertization of hazardous dredging spoils. Waste Manag 22:865–869
86. Brusatin G, Bernardo E, Andreola F, Barbieri L, Lancellotti I, Hreglich S (2005) Reutilization of waste inert glass from the disposal of polluted dredging spoils by the obtainment of ceramic products for tiles applications. J Mater Sci 40:5259–5264
87. Suzuki S, Tanaka M, Kaneko T (1997) Glass-ceramic from sewage sludge ash. J Mater Sci 32:1775–1779
88. Park YJ, Moon So, Heo J (2003) Crystalline phase control of glass ceramics obtained from sewage sludge fly ash. Ceram Int 29:223–227
89. Toya T, Nakamura A, Kameshima Y, Nakajima A, Okada K (2007) Glass-ceramics prepared from sludge generated by a water purification plant. Ceram Int 33:573–577
90. Bhat PN, Ghosh DK, Desai MVM (2002) Immobilisation of beryllium in solid waste (red-mud) by fixation and vitrification. Waste Manag 22:549–556
91. Roberts D, Stuart JH (1989) Vitrification of asbestos waste. US Patent 4,820,328, 11 April 1989
92. Dall'Igna R, Falcone R, Hreglich S, Profilo B, Vallotto M, Cadore A, Grattieri W (2002) Production of mineral fertilizer glass from inertized asbestos containing waste. Riv Staz Sper Vetro 6:13–15
93. Bernardo E, Scarinci G, Hreglich S (2005) Foam glass as a way of recycling glasses from cathode ray tubes. Glass Sci Technol 78:7–11
94. Bernardo E, Cedro R, Florean M, Hreglich S (2007) Reutilization and stabilization of wastes by the production of glass foams. Ceram Int 33:963–968
95. Bernardo E, Scarinci G, Bertuzzi P, Ercole P, Ramon L (2009) Recycling of waste glasses into glass and glass-ceramic foams. J Porous Mater 17(3):359–365
96. Bernardo E, Andreola F, Barbieri L, Lancellotti I (2005) Sintered glass-ceramics and glass-ceramic matrix composites from CRT panel glass. J Am Ceram Soc 88:1886–1891
97. Bernardo E, Castellan R, Hreglich S, Lancellotti I (2006) Sintered sanidine glass-ceramics from industrial wastes. J Eur Ceram Soc 26:3335–3341
98. Tucci A, Esposito L, Rastelli E, Palmonari C, Rambaldi E (2004) Use of soda-lime scrap-glass as a fluxing agent in a porcelain stoneware tile mix. J Eur Ceram Soc 24:83–92
99. Pontikes Y, Christogerou A, Angelopoulos G, Rambaldi E, Esposito L, Tucci A (2005) Use of soda-lime-silica scrap glass in the traditional ceramic industry. Glass Technol 46:200–207
100. Tarvornpanich T, Souza GP, Lee WE (2005) Microstructural evolution on firing soda-lime-silica glass fluxed whitewares. J Am Ceram Soc 88:1302–1308
101. Tucci A, Rambaldi E, Esposito L (2006) Use of scrap glass as raw materials for porcelain stoneware tiles. Adv Appl Ceram 105:40–45
102. Raimondo M, Zanelli C, Matteucci F, Guarini G, Dondi M, Labrincha JA (2007) Effect of waste glass (TV/PC cathodic tube and screen) on technological properties and sintering behaviour of porcelain stoneware tiles. Ceram Int 33:615–623
103. Rawlings RD, Wu JP, Boccaccini AR (2006) Glass-ceramics: their production from wastes – a review. J Mater Sci 41:733–761
104. Bernardo E, Esposito L, Rambaldi E, Tucci A (2009) Glass-based stoneware as a promising route for the recycling of waste glasses. Adv Appl Ceram 108:2–8
105. http://www.europlasma.com/
106. Morimoto N et al (1988) Nomenclature of pyroxenes. Am Mineralog 73:1123–1133
107. Peng F, Liang K, Hu A (2005) Nano-crystal glass–ceramics obtained from high alumina coal fly ash. Fuel 84:341–346
108. Karamanov A, Cantalini C, Pelino M, Hreglich S (1999) Kinetics of phase formation in jarosite glass-ceramic. J Eur Ceram Soc 19:527–533

109. Karamanov A, Pelino M (2001) Crystallization phenomena in iron-rich glasses. J Noncryst Solids 281:139–151
110. Bloomer PE, Feng X, Chantaraprachoom N, Gong M, McCready DE (1999) Effect of crystallization, redox, and waste loading on the properties of several glassy waste forms. J Am Ceram Soc 11:2999–3011
111. Gutzow I, Pascova R, Karamanov A, Schmelzer J (1998) The kinetics of surface induced sinter-crystallization and the formation of glass-ceramic materials. J Mater Sci 33:5265–5273
112. Müller R, Zanotto ED, Fokin VM (2000) Surface crystallization of silicate glasses: nucleation sites and kinetics. J Noncryst Solids 274:208–231
113. Prado MO, Zanotto ED (2002) Glass sintering with concurrent crystallization. C R Chimie 5:773–786
114. Francis AA, Rawlings RD, Sweeney R, Boccaccini AR (2004) Crystallization kinetic of glass particles prepared from a mixture of coal ash and soda-lime cullet glass. J Noncryst Solids 333:187–193
115. Hernandez-Crespo MaS, Romero M, Rincon JMa (2006) Nucleation and crystal growth of glasses produced by a generic plasma arc-process. J Eur Ceram Soc 26:1679–1685
116. Bernardo E (2008) Fast Sinter-crystallization of a glass from waste materials. J Noncryst Solids 354:3486–3490
117. Karamanov A, Pisciella P, Cantalini C, Pelino M (2000) Influence of Fe^{3+}/Fe^{2+} ratio on the crystallization of iron-rich glasses made with industrial waste. J Am Ceram Soc 83:3153–3157
118. Karamanov A, Pelino M, Hreglich S (2003) Sintered glass-ceramics from municipal solid waste-incinerator fly ashes-part I: the influence of the heating rate on the sinter-crystallization. J Eur Ceram Soc 23:827–832
119. Karamanov A, Aloisi M, Pelino M (2005) Sintering behaviour of a glass obtained from MSWI ash. J Eur Ceram Soc 25:1531–1540
120. Ray A, Tiwari AN (2001) Compaction and sintering behaviour of glass-alumina composites. Mater Chem Phys 67:220–225
121. Romero M, Rincon JMa (1999) Surface and bulk crystallization of glass-ceramic in the Na_2O-CaO-ZnO-PbO-Fe_2O_3-Al_2O_3-SiO_2 system derived from a goethite waste. J Am Ceram Soc 82:1313–1317
122. Francis AA, Rawlings RD, Sweeney R, Boccaccini AR (2002) Processing of coal ash into glass ceramic products by powder technology and sintering. Glass Technol 43:58–62
123. Fidancevska E, Mangutova B, Milosevski D, Milosevski M, Bossert J (2003) Sci Sinter 35:85–91
124. Scarinci G, Brusatin G, Barbieri L, Corradi A, Lancellotti I, Colombo P, Hreglich S, Dall'Igna R (2000) Vitrification of industrial and natural waste with production of glass fibres. J Eur Ceram Soc 20:2485–2490
125. Hreglich S, Cioffi F (2009) Continuous glass fibres from waste and their application in reinforced materials. Adv Appl Ceram 108:22–26
126. Hreglich S, Falcone R, Vallotto M (2001) The recycling of end of life panel glass from TV sets in glass fibres and ceramic productions. In: Dhir RK, Limbachiya MC, Dyer TD (eds) Recycling and reuse of glass cullet. Thomas Telford, London, pp 123–134
127. Marabini AM, Plescia P, Maccari D, Burragato F, Pelino M (1998) New materials from industrial and mining waste: glass-ceramics and glass- and rock-wool fibre. Int J Miner Process 53:121–134
128. Scarinci G, Brusatin G, Bernardo E (2005) Production technology of glass foams. In: Scheffler M, Colombo P (eds) Cellular ceramics. structure, manufacturing, properties and applications. Wiley-VCH, Weinheim
129. Méar F, Yot P, Viennois R, Ribes M (2007) Mechanical behaviour and thermal and electrical properties of foam glass. Ceram Int 33:543–550
130. Brusatin G, Scarinci G, Zampieri L, Colombo P (2002) Foam glass from cullet. Glass Mach Plant Accessory 1:108–110
131. Fernandes HR, Tulyaganov DU, Ferreira JMF (2009) Production and characterisation of glass ceramic foams from recycled raw materials. Adv Appl Ceram 108:9–13
132. Tulyaganov DU, Fernandes HR, Agathopoulos S, Ferreira JMF (2006) Preparation and characterization of high compressive strength foams from sheet glass. J Porous Mater 13:133–139
133. Wu JP, Boccaccini AR, Lee PD, Kershaw MJ, Rawlings RD (2006) Glass ceramic foams from coal ash and waste glass: Production and characterisation. Adv Appl Ceram 105:32–39
134. Bernardo E (2007) Micro- and macro-cellular sintered glass-ceramics from wastes. J Eur Ceram Soc 27:2415–2422
135. Boccaccini AR, Bücker M, Bossert J, Marszalek K (1997) Glass matrix composites from coal fly ash and waste glass. Waste Manag 17:39–45
136. Ferraris M, Salvo M, Smeacetto F, Augier L, Barbieri L, Corradi A, Lancellotti I (2001) Glass matrix composites from solid waste materials. J Eur Ceram Soc 21:453–460
137. Appendino P, Ferraris M, Matekovits I, Salvo M (2004) Production of glass-ceramic bodies from the bottom ashes of municipal solid waste incinerators. J Eur Ceram Soc 24:803–810
138. Aloisi M, Karamanov A, Taglieri G, Ferrante F, Pelino M (2006) Sintered glass ceramic composites from vitrified municipal solid waste bottom ashes. J Hazard Mater 137:138–143
139. Bernardo E, Scarinci G, Hreglich S (2005) Development and mechanical characterization of Al_2O_3 platelet-reinforced glass matrix composites obtained from glasses coming from dismantled cathode ray tubes. J Eur Ceram Soc 255:1541–1550
140. Bernardo E, Castellan R, Hreglich S (2007) Al_2O_3-platelet reinforced glass matrix composites from a mixture of wastes. J Mater Sci 42:2706–2711
141. Rozenstrauha I, Cimdins R, Berzina L, Bajare D, Bossert J, Boccaccini AR (2002) Sintered glass-ceramic matrix composites made from Latvian silicate wastes. Glass Sci Technol 75:132–139
142. Saccani A, Sandrolini F, Barbieri L, Corradi A, Lancellotti I (2001) Structural studies and electrical properties of recycled glasses from glass and incinerator waste. J Mater Sci 36: 2173–2177
143. Barba MF, Callejas P, Arabe JO, Ajò D (1998) Characterization of two frit ceramics materials in low cost fertilizers. J Eur Ceram Soc 18:1313–1317
144. Jin W, Meyer C, Baxter S (2000) Glascrete-concrete with glass aggregate. ACI Mater J 97:208–213

145. Schroeder RL (1994) The use of recycled materials in highway construction. Public Roads 58:32–41
146. Su N, Chen JS (2002) Engineering properties of asphalt concrete made with recycled glass. Resour Conserv Recycl 35:259–274
147. Gao Z, Drummond CH (1999) Thermal analysis of nucleation and growth of crystalline phases in vitrified industrial waste. J Am Ceram Soc 82:561–565
148. Romero M, Rincon JMa, Acosta A (2002) Effect of iron oxide content on the crystallisation of a diopside glass-ceramic glaze. J Eur Ceram Soc 22:883–890
149. Zubekhin AP, Zhabrev VA, Kondyurin AM (1993) Glass formation and crystallization in the SiO_2–CaO–MgO–Fe_2O_3–MnO_2–K_2O–Na_2O for synthesizing heat resistant coatings. Steklo i Keramica 5:26–28
150. Barbieri L, Corradi A, Lancellotti I, Manfredini T (2002) Use of municipal incinerator bottom ash as sintering promoter. Waste Manag 22:859–863
151. Bernardo E, Esposito L, Rambaldi E, Tucci A, Hreglich S (2008) Recycle of waste glass into "Glass-ceramic Stoneware". J Am Ceram Soc 91:2156–2162

Books and Reviews

Gomez E, Rani DA, Cheeseman CR, Deegan D, Wise M, Boccaccini AR (2008) Thermal plasma technology for the treatment of wastes: a critical review. J Hazard Mater 161:614–626
Oh CO (2001) Hazardous and radioactive waste treatment technology. CRC Press, Boca Raton
Scholze H (1991) Glass: nature, structure and properties. Springer, New York
Strnad Z (1986) Glass-ceramic materials. Elsevier, Amsterdam
Vesilind PA, Worrell W, Reinhart D (2002) Solid waste engineering. Rooks/Cole, Pacific Grove

Volcanoes of Mexico

Nick Varley
Centre of Exchange and Research in Volcanology, Universidad de Colima, Col. Villas San Sebastián, Colima, Mexico

Article Outline

Glossary

Andesite Magma of intermediate composition of silica, a common product of eruptions from stratovolcanoes.

Basalt Magma of a low silica composition, has a low viscosity and usually is emplaced in lava flows or as scoria.

Dacite Magma of a composition higher in silica, between andesite and rhyolite.

Debris avalanche Large collapse event which produces an extensive deposit, often characterized by hummock-shaped hillocks. Occur from most likely all stratovolcanoes, due to their inherent instability.

Holocene The time period since the end of the last glaciation (11,700 years ago to present).

Lahar Mudflow with the remobilization of active volcanic pyroclastic deposits. Can be hot if occurring during or soon after the eruption.

Maar Explosive craters surrounded by a ring-shaped deposit of pyroclasts. Form due to phreatomagmatic interactions, often in basins with extensive aquifers.

Monogenetic field Region featuring the growth of cinder cones or maars, with each eruptive centre typically exhibiting only one eruption.

Phreatomagmatic Explosive eruption resulting from the interaction between magma and water. Results in the expulsion of juvenile material. *Phreatic* implies no juvenile component.

Pleistocene The epoch between 2.58 million and 11,700 years ago, which included the major period of glaciation.

Pyroclastic flow or pyroclastic density current A gravity-driven mobile flow of hot gases mixed with ash and rocks, which descend volcano flanks

at fast speeds posing a great threat to anything in their path.

Rhyolite Evolved magma with a high silica contents, often associated with large explosive eruptions.

Strombolian Explosive eruption of low magnitude, occur usually with low-viscosity basaltic magma. Activity that produces scoria cones: accumulation of vesiculated magma around the vent.

Surtseyan type Eruption occurring through shallow water with the formation of an island from the deposition of pyroclastic material resulting from the highly explosive interaction between magma and seawater.

Vulcanian Explosive eruption of medium magnitude: result from the failure of an impermeable layer above the magma column, which restricts the process of magma degassing. Produce a lot of ballistics.

Xenolith Material of external origin and different composition incorporated within magma. They are often pieces of upper mantle or lower crustal rocks.

Definition of the Subject

Volcanoes represent one of nature's most formidable yet beautiful spectacles. They represent an omnipresent threat in many parts of the world, but also attract an increasing number of visitors, who have the urge to scale their flanks and peer into the depths of their craters. This entry includes a brief introduction to volcanoes: the reason they form, where they are located, and the hazard presented by the products of different types of eruption. Mexico is one of the world's most volcanic regions and a summary of the volcanoes of this country is presented. Included is a list of the active volcanoes of Mexico, defined for the first time using systematic criteria.

Introduction

In a small village somewhere in Mexico, a farmer awoke one morning and headed off to tend his crops, not expecting that anything out of the ordinary was going to happen that day. But his farm included a field that was not any ordinary cornfield, but a cornfield which all Mexican children now learn about in primary school. It was a cornfield which on that particular day was to become host to the birth a volcano. That farmer was to become one of the few people to witness a birth of one of nature's most spectacular creations.

The eruption of Paricutín took place between 1943 and 1952. It is the youngest of at least 1,040 volcanoes that are located in the Michoacán-Guanajuato volcanic field. Perhaps there is no better country than Mexico to study this type of volcanism. The country hosts at least 13 major monogenetic fields, each with countless scoria cones and other features. Many new fields are only now being defined.

The Basics of Volcanism

Volcanoes are located within certain regions on the Earth's surface related either to plate tectonics or to so-called hot spots, which represent locations below which convection-driven mantle plumes carry hotter material toward the surface. Plate boundaries can be divergent, such as mid-ocean ridges, which host the largest numbers of volcanoes on the planet, or rift zones, such as in East Africa. Other boundaries are convergent, which results in subduction, if one or both plates are thinner oceanic plates. In this process, one plate descends below the other carrying ample quantities of water within the rocks' structure. The water is a key ingredient to the formation of volcanoes since it lowers the melting point of the mantle, which becomes buoyant and rises toward the surface. A variable amount of the subducted plate is carried with it. Rather than erupt at the surface, the majority of magma accumulates within the crust after it loses its buoyancy, then cools with crystals slowly forming to create intrusive igneous rock. Some magma however will make it to the surface producing an eruption.

Volcanic eruptions can be divided into two broad categories: explosive or effusive. Various factors combine to determine how the magma emerges, the key ingredient again is water. Magma contains typically a few percent of water by weight, which is adequate to cause violent explosive eruptions if the magma rises to the surface sufficiently quickly. During ascent, decompression means that the dissolved water (and other gases) starts to form bubbles, which grow, and in the case of explosive eruptions, the volume of gas becomes much larger than the residual liquid magma. This gas can expand at alarming rate, which can produce the largest eruptions known as Plinian. Plinian eruptions

produce columns of gas and ash which can rise to 20 km or more above the surface. The larger eruptions produce enormous clouds of ash which enter the stratosphere, are dispersed widely by winds, and promote significant reductions in atmospheric temperatures. If the volume of magma emerging is large and its ascent rapid, the evacuation of the magma chamber below can promote collapse and the formation of a caldera. Some historical caldera-forming eruptions have had a considerable impact on the world's climate (e.g., the volcano Tambora, Indonesia, in 1815).

Less explosive types of explosive eruption, such as Strombolian, result from the ascent of less viscous magma (such as basalt). Bubbles of volcanic gas are able to flow through the magma, which prevents the buildup of extremely high pressures. This type of activity can result in the construction of a scoria cone, a common feature of the Mexican landscape. Larger stratovolcanoes are constructed over thousands of years with the erupted magma accumulating to form the volcanic edifice. The interaction of the rising magma with the ocean, a lake, or groundwater can escalate the explosivity, through the formation of rapidly expanding steam, and produce phreatic or phreatomagmatic eruptions. On land, the result can be a maar.

Effusive eruptions on the other hand produce lava domes and flows. In this case, the most important characteristics of the ascending magma: its temperature, viscosity, volatile contents, amount of crystals, combine to produce this less hazardous type of eruption. Lava flows can displace human settlements and destroy cultivated land, such as what happened at Paricutín, but they seldom endanger lives.

The products of explosive eruptions consist of ballistic rocks, ash, and pyroclastic density currents. The secondary remobilization of the pyroclastic deposits by water can produce immense mudflows or lahars. With the exception of ballistics, which do not reach very far from the volcano, each of these hazards can present a major risk to the population living close-by. Volcanologists need to investigate the deposits of previous eruptions to understand what might happen in the future. Hazard maps can be created to indicate where the eruptive products might accumulate. Several of Mexico's volcanoes have hazard maps and for several others, maps are in progress.

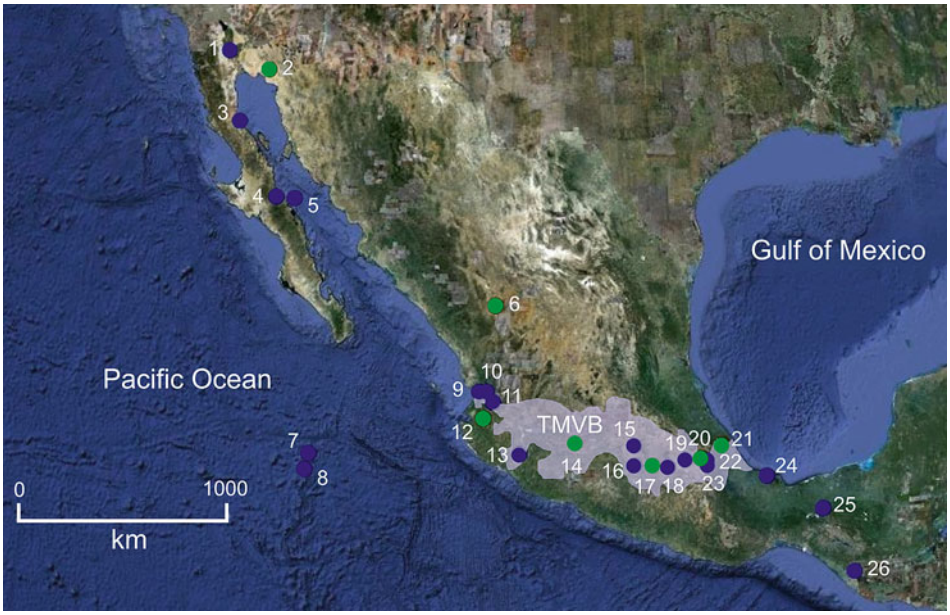
The Importance of Mexico's Volcanoes

Figure 1 shows the location of the major volcanoes of Mexico. The majority are located in the Trans-Mexican Volcanic Belt (TMVB), which extends across the country and has been described as the world's largest intracontinental volcanic arc. Over 40% of the population lives in this zone, which includes the large cities of Mexico City, Guadalajara, and Puebla. This makes volcanic risk an important issue. Enormous stratovolcanoes lie here within, like Nevado de Toluca, Popocatépetl, and the tallest volcano in North America: Citlaltépetl (or Pico de Orizaba, the name given by the Spanish colonialists). Construction of these edifices took place from the late Pleistocene onward. Figure 2 illustrates the magnitude of the most important eruptions of Mexico's active volcanoes.

Apart from the eruption of magma, stratovolcanoes can drastically influence their surroundings in another form: debris avalanches. These are huge collapses of a large proportion of the volcanic edifice and are common within the lifetime of a large volcano, given its unstable nature. Mexico has many examples with extensive deposits [1], which represent an important contribution to the shaping of the landscape.

The TMVB also includes nine identified large calderas with their associated ignimbrite deposits, such as Los Humeros or La Primavera, with 182 other circular features identified on satellite images [2]. Most likely many of these are also collapse caldera structures. It is easy to observe other expressions of volcanism such as maars, shield volcanoes, fissure eruptions, domes, volcanic islands, etc. Some of these calderas play an important role in the energy supply in Mexico. Currently some 958 MW of geothermal power are produced, which puts the country in fourth place in the world's rankings [3]. Production is dominated by Cerro Prieto in Baja California, with Los Azufres and Los Humeros contributing in the TMVB.

Volcanoes make their appearance throughout the different chapters of Mexican history. Eruptions of Popocatépetl drove the population out of prehispanic cities such as Cuicuilco, on the outskirts of Mexico City. The Aztecs reported various eruptions of Popocatépetl, which means "the smoking mountain" in the Náhuatl language. Other sacred centres, such as La



Volcanoes of Mexico. Figure 1

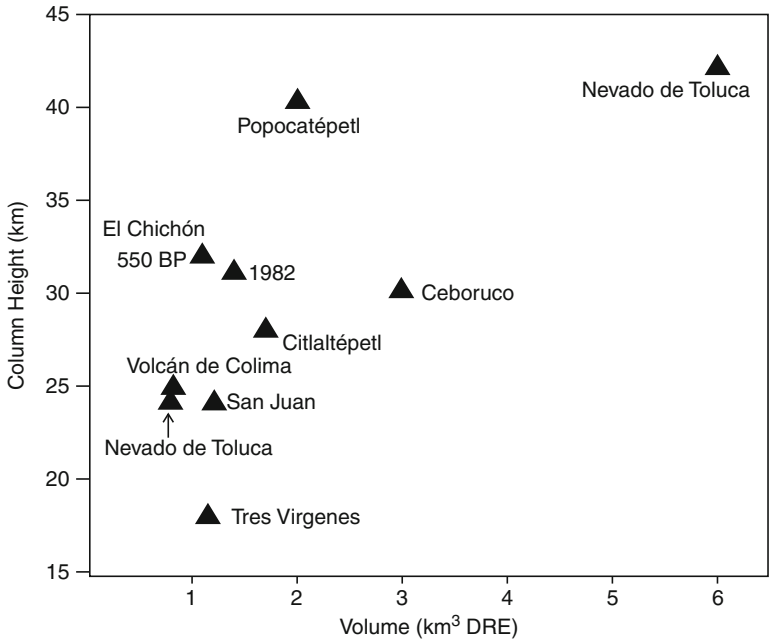
Map of Mexico showing the locations of the active volcanoes or volcanic fields. TMVB: Trans-Mexican Volcanic Belt. *1* – Cerro Prieto; *2* – Pinacate; *3* – Isla Tortuga; *4* – Tres Virgenes; *5* – Isla San Luis; *6* – Durango Volcano Field; *7* – Bárcena; *8* – Isla Socorro; *9* – San Juan; *10* – Sangangüey; *11* – Ceboruco; *12* – Mascota Volcanic Field; *13* – Volcán de Colima; *14* – Michoacán-Guanajuato Volcanic Field; *15* – Jocotitlán; *16* – Nevado de Toluca; *17* – Chichinautzin Volcanic Field; *18* – Popocatépetl; *19* – La Malinche; *20* – Serdán-Oriental volcanic field; *21* – Naolinco volcanic field; *22* – Las Cumbres; *23* – Citlaltépetl; *24* – San Martín; *25* – El Chichón; *26* – Tacaná

Campana in Colima, show evidence that the construction of certain pyramids tried to mimic distant volcanic peaks. Vast valleys are filled with soils that owe their fertility to the outpourings of countless eruptions, and other value has been reaped, such as the hugely important volcanic product that is obsidian. This volcanic glass, usually shiny black in colour, can be fashioned without much skill into knifes and weapons. It was clearly important in the development of the early Mexican people. Cortes during his conquest of Mexico had run out of gunpowder. Unfortunately for countless indigenous people, he was able to get the sulphur to replenish his stocks by sending his men into the crater at Popocatépetl. The Paso de Cortes, situated high up between Popocatépetl and its older neighbour Iztaccíhuatl, is host to the only statue to Cortes in the whole of Mexico.

During the nineteenth century various pioneers of exploration and scientific observation made important discoveries, and their documents along with the paintings of various travellers are important documentation of the early eruptive record of many volcanoes.

Volcanology in Mexico is a young science, having only taken off following the 1982 eruption of El Chichón. An enormous amount of work remains to be done: identification of deposits; geochemical analyses to determine details of magma ascent and emplacement; geophysical work to establish the location of magma chambers, hydrothermal systems, etc.; risk assessments and the creation of hazard maps. There are countless young lava flows, some possibly matched by accounts of activity from prehispanic stories or texts from European missionaries or explorers. But many question marks remain regarding the date of the most recent activity for many centres.

The number of volcanoes described as being "active" varies depending upon the text being consulted. Here, the list in Table 1 includes 26 volcanoes or volcanic fields ascribed as being active, and represents the first systematic attempt to create



Volcanoes of Mexico. Figure 2

Eruption column height v. erupted volume of magma for Mexico's most recent large (Plinian) eruptions. DRE = Dense rock equivalent, thereby taking into account any vesicles or pores within the deposits. Eruption ages: El Chichón 1982 and 550 year BP; Volcán de Colima 1913; Ceboruco 1,060 ± 55 year BP; Popocatépetl 4,965 ± 65 year BP; Tres Vírgenes 6,500 year BP; Citlaltépetl 8,500–9,000 year BP; Nevado de Toluca 10,500 year BP and 21,700 year BP; San Juan 14,770 ± 480 year BP (Figure modified from Arana-Salinas et al. 2010)

a definitive list. An active volcano is one that could erupt again, which means the presence of a magma body present at depth. This should give certain telltale signs, such as seismicity, heat dissipated through the interaction with ground water, resulting in hot springs, or direct degassing in the form of fumaroles. A seismic network may not be present to detect these movements and the thermal manifestations will not always be present. Medina et al. [4] in their early work considered the existence of fumarolic activity as a criterion for considering a volcano active. The other consideration is the repose interval since the last eruption. Mexico has many volcanoes with long repose intervals, often of several thousand years, hence a somewhat arbitrary limit can be taken as 10,000 years (following the definition of the Smithsonian Institute, Washington, D.C.). Since volcanic fields feature monogenetic volcanoes (those that erupt just once), these are taken as a single active entity. Several other volcanic fields, or scoria cone complexes not considered here, could also be active, in that they feature lava emissions that may have occurred within the Holocene. However, they have not been included in the list since no strong evidence exists to date. Examples are La Gloria field in Veracruz and Isla Isabel in Nayarit, where, as is the case at many locations, further studies are needed to clarify the situation. Three Pleistocene calderas in the TMVB still have fumarolic activity with extensive geothermal fields: La Primavera, Los Azufres, and Los Humeros. These have not been included as active since the probability of future eruptive activity is regarded as negligible, given the long periods that have passed since their last eruptions. Prehistoric calderas sometimes feature persistent hydrothermal activity, which unlike the case of stratovolcanoes or domes, is not necessarily regarded as evidence for the likelihood of a future eruption.

Monitoring has been expanded greatly in the last 20 years with sophisticated networks feeling the pulse of Popocatépetl and Volcán de Colima. Seismic and geochemical monitoring (water samples) is being carried out at El Chichón and Tacaná with some seismic

Volcanoes of Mexico. Table 1 Active volcanoes of Mexico, in order starting with the most recently active

Volcano/field	Date of last eruption	Details	References
Volcán de Colima	Currently erupting[a]	Cyclic effusive and Vulcanian explosions	[38]
Popocatépetl	Currently erupting[a]	Cyclic effusive and Vulcanian explosions	[42]
Isla Socorro	1993	Submarine flank eruption	[21]
Tacaná	1986	Phreatic explosion	[67]
El Chichón	1982	Plinian eruption	[59]
Bárcena	1953	Explosive eruption	[20]
Michoacán-Guanajuato volcanic field	1943–1952	Formation of cinder cone – Paricutín	[50]
Ceboruco	1870–1875	Dacitic lava flow	[26]
Citlaltépetl	1846	Explosive	[47]
St. Martin Tuxtla	1793	Summit cinder cones	[58]
Tres Virgenes	1746	Explosive though unconfirmed	[15]
Sangangüey	1742	Unconfirmed eruption of flank cone	[27]
Jocotitlán	1270	Explosive	[41]
Chichinautzin volcanic field	340 AD	Xitle cinder cone and flow	[55]
La Malinche	1170 BC	Ash fall and pyroclastic flow	[45]
Naolinco volcanic field	1200 BC	Cinder cone and lava flow	[57]
Isla San Luis	Recent but undefined age	Rhyolitic obsidian domes	[16]
Nevado de Toluca	3,300 BP	Pyroclastic flow and surge	[40]
Mascota Volcanic Field	Maybe 5,600 BP	Lava flow and scoria	[53]
Las Cumbres	5,900 BP	Rhyolitic lava dome	[49]
Durango Volcanic field	Few thousand years	Maar formation, scoria cones and lava flows	[52]
Pinacate volcanic field	Holocene	Unknown	[10]
Cerro Prieto	Holocene	Unknown	[3]
Isla Tortuga	Holocene	Unknown	[17]
Serdán-Oriental volcanic field	Holocene or late Pleistocene	Las Derrumbadas – fumaroles	[56]
San Juan	Holocene or late Pleistocene	Plinian eruption 14,770 BP	[28]

[a]Effusive eruption still ongoing at the time of writing: June 2011

monitoring of Citlaltépetl and Ceboruco. Much effort is being directed at increasing the quality and diversity of the information being generated, as well as the definition of models of the eruption mechanisms, thus advancing what can be interpreted from the data.

Tectonics

Volcanism in the north of Mexico can be clearly divided into various regions, each with its characteristic tectonic situation. Volcanism can be associated with

extension; this gives rise to various volcanic centres in Baja California and Sonora. Further south the oceanic Rivera and Cocos plates collide with the continental North America plate; the resulting subduction gives rise to the majority of volcanoes which are located in the TMVB. Further to the south the Central American Volcanic Belt (CAVB) starts with the potentially dangerous volcano Tacaná, shared between Mexico and Guatemala. Whereas this belt, which extends through Central America, is parallel to the subduction trench, the TMVB clearly extends obliquely across the country, making an angle of about 15° with respect to the Middle American Trench. This intriguing characteristic has been explained using geophysical and geochemical observations that suggest a decreasing angle of subduction as one moves down the coast from the NW to SE [5, 6]. This results in an increasing distance between the trench, the point where the descent of the oceanic plate commences, and the zone of melting, which typically occurs at pressures corresponding to a 100 km depth.

One fascinating feature within the belt is the existence of several small chains of volcanoes, which demonstrate a decreasing age from north to south: Cántaro – Nevado de Colima – Volcán de Colima; Tlaloc – Itzaccihuatl – Popocatépetl; and Cofre de Perote – Las Cumbres – Citlaltépetl –Sierra Negra. Evidence of an overall southern migration of the magma chamber has been identified [7].

At each end of the TMVB things get tectonically more complicated. Firstly to the east, between the TMVB and CAVB the very active volcano El Chichón is found along with other extinct peaks forming the Pliocene to Recent age Chiapanecan Volcanic Arc (CVA). This rather esoteric arc stretches some 150 km in a NW to SE direction. The subducting Cocos plate changes from being a virtually flat slab in Central Mexico to a ∼45° dip angle beneath the CVA, with the slab being located an unusually deep ∼200 km below the arc [8]. At the north-western extreme of the TMVB evidence of rifting, or the separation of two plates, is present. Here magmas with a higher alkaline content have been erupted, which combined with geomorphological evidence, such as the alignment of scoria cones or fault scarps, suggests that the so-called Jalisco Block might one day separate from mainland Mexico [9, 10].

The following summary of Mexico's different volcanic centres starts in the NW of the country, continues with the TMVB, going from W to E, continues with monogenetic fields and finishes with the volcanoes of Chiapas. The account concentrates on active volcanic centres, though some important extinct volcanoes are included.

Baja California and Sonora

Sierra Pinacate Volcanic Field

This region includes more than 400 cinder cones and lava flows, and eight large maar craters of late-Pleistocene to Holocene age [11]. As with many cases of volcanic fields in Mexico, the cones follow alignments determined by the regional tectonics. Maars are the result of phreatomagmatic activity and this field has excellent examples. Crater Elegante is the largest with a diameter of 1.6 km and depth of 240 m. Eruptions here generally commenced with Stombolian eruptions and the construction of cinder cones. Unusually, the phreatomagmatism occurred later, probably a reflection of the aridity of the region. Despite a lack of any firm dates of Pinacate rocks placing the most recent activity from the field in the Holocene, the low level of erosion suggests activity within this period and hence its inclusion in the list of active Mexican volcanic fields [3].

Cerro Prieto

This is the location of the most productive geothermal field in Mexico. The dacitic lava dome is located within an active continental rift, which marks the transition from the famous San Andreas transform fault system to the north to the spreading ridge of the East Pacific Rise in the Gulf of California to the south. Geophysical evidence gave 10,000 years as an approximate youngest date [12] and the presence of geothermal manifestations at the surface also suggests its inclusion in the active list. In addition, legends of the local Cucupas people describe hot rocks being thrown by a monster and fires coming from the soil, suggesting it may have been active much more recently [4].

San Quintín, Jaraguay, and San Borja Volcanic Fields

These three fields are not being included as active, although some early authors thought they are of Holocene age [4, 13], there is no clear evidence. Recent

dating suggests the San Quíntin field is no younger than 20,000 years [14]; however, the Jaraguay and San Borja fields have less vegetation on their flows, suggesting younger ages [15]. The three fields consist of cinder cones and lava flows, San Quintín with ten distinct complexes [13].

Tres Virgenes

The only large stratovolcano in the Baja California region is the Tres Virgenes complex. It consists of three cones, with a progression to younger ages to the SW, La Virgen being the youngest. There may have been an eruption in 1746, according to a record of observations, but there is no hard evidence [16]. The last major Plinian eruption occurred about 6,500 years ago. Volcanism here is associated with dominantly extensional faulting.

Isla San Luis and Isla Tortuga

These two islands represent the youngest volcanism in the Sea of Cortés. The oldest deposits of Isla San Luis show that it was born like the island of Surtsey, in Iceland, with corresponding highly explosive eruptions producing pyroclastic surges [17]. This activity was followed by dacitic flows and the formation of tuff rings, then finally two rhyolitic domes were emplaced. The activity can be thus characterized by a successive eruption of progressively more differentiated lavas. Suggestions have been made that the two rhyolitic domes are less than 100 years old [4]. While this might be wishful thinking, they certainly are relatively young, warranting the inclusion of this island in the active volcanoes list.

Isla Tortuga is a shield volcano with young lava flows located further south. The latest stage of activity culminated in caldera collapse, extrusion of the surficial flows, and the formation of a lava lake. Medina et al. [4] defined it as being Holocene, though no evidence was given. The spatter cones within the caldera appear to be recent as do many of the flows that cover its flanks [18].

Isla Isabel

This small island off the state of Nayarit represents an emergent Surtseyan-type volcanic complex. Various craters on this island show evidence that there was a general migration of volcanic activity from northwest to southeast [19]. It is one locality where it is possible to find abundant mantle xenoliths. It has been suggested that the most recent eruptions were less than 10,000 years ago, though due to the lack of hard evidence, the island is not included in the list of active Mexican volcanoes.

Pacific Islands

The Revillagigedo archipelago is volcanic with two active centres. It is located on the Mathematician Ridge, which was the location of an active spreading centre prior to its migration 3.15 Ma BP [20]. The volcano Bárcena on the small 4.5 km long unpopulated island of San Benedicto formed in an eruption in 1952–1953 [21]. Since, it has remained quiet and unstudied.

Famous for its marine life and endemic species of fauna and flora, Isla Socorro represents the other active volcano of the archipelago. The only human occupancy is a naval base in the SE corner of the island. In Fig. 3 you can see a view from off the southern coast, showing domes and old lava flows, one with beautiful levees (ridges at the edge of the flow parallel to its flow direction). The last eruption occurred in 1993: a relatively small submarine event which sent blocks of gas-filled pumice hurtling to the surface [22]. The summit area has not seen an eruption since at least 15,000 years ago [20], but has an extensive hydrothermal field, emitting gases which contain unusually high concentrations of methane and hydrogen [23]. The island is covered with extensive lava flows and domes, unusually of peralkaline composition, and more recent cinder cones in the SE part of the island. Here the ^{14}C method was used to date some lacustrine deposits, which are clearly older than nearby cones [24]. An age of about 5,000 years was obtained.

Tepic-Zacoalco Rift

This zone stretches from beyond Tepic southeast to Guadalajara. It includes a mesmerizing density of volcanic features, including five andesitic stratovolcanoes, plus maars, domes, and countless scoria cones. Many groups of cinder cones are aligned in a NW-SE direction, the same as the rift. To the NE of

Volcanoes of Mexico. Figure 3
View from the south side of Isla Socorro. Various domes and lava flows can be distinguished on the slopes of the volcano

the zone the mountains represent the huge deposits of the Sierra Madre Occidental volcanic province, while to the SW the micro-plate known as the Jalisco Block is located [25]. The zone is dominated by several large stratovolcanoes, but also includes many fascinating features such as the maar Santa Maria del Oro, which contains a picturesque lake.

Ceboruco

The most recent activity in this region was a lava flow which emanated from Ceboruco in 1870 and descended down the south-west flank. Recent activity at this volcano has been in the form of various lava flows of either a dacitic or andesitic composition, some originating from near the summit, others from lower on the flanks. Two roads extend along the rift zone, one to the north, the other to the south of Ceboruco. Both provide stunning views of the many lava flows which have covered the flanks of this volcano. The last major explosive eruption occurred only 1,000 years ago [26] and produced an extensive pyroclastic flow and pumice fall deposit. A caldera was formed with a 4 km diameter. The eruptive episode was terminated by the emplacement of two small domes and a lava flow within the crater. Following, there was a period of about 500 years with the formation of six different andesitic or dacitic lava flows on the flanks [27]. Since then the volcano has been quieter, with just the one recent 7 km long lava flow.

Sangangüey and San Juan

Further to the NW, Sangangüey is an impressive mass with a spectacular lava spine at its summit and large collapse scars on its flanks. The spine can be seen in Fig. 4, which was taken from the summit of Ceboruco. Many cinder cones exist along the rift, all aligned in the NW-SE direction in five different linear groups. There are no records of historic activity from the peak, although historical accounts coupled with low levels of erosion suggest that some of the flank cinder cones have erupted less than 1,000 years ago [28]. For this reason the volcano is included in the list of active volcanoes of Mexico.

At the Tepic end of the rift, in fact adjacent to the city of Tepic, there lies San Juan, a volcano which is not termed active using the 10,000 year rule. However, it has been pointed out that its explosive past, with the last Plinian event occurring about 14,770 years ago, suggests that a reawakening should not be totally ruled out [29].

Volcanoes of Mexico. Figure 4
Sangangüey shown in the distance, looking from Ceboruco. Many scoria cones can be seen following the NW-SE trending faults

Volcán Tequila

Tequila, apart from being a notorious Mexican export, gives its name to a large stratovolcano, located further to the SE along the rift. The youngest associated volcanic feature (~60 ka) is a small andesitic vent called Cerro Tomasillo [30]. The region features various cones, domes, and flows with a large variety of compositions from basalt to rhyolite. Volcán Tequila itself also has a summit spine; in this case it is a 300 m high pinnacle, which dominates any view to the volcano from an east or north direction.

La Primavera

Very close to Mexico's second largest city, Guadalajara, there is a picturesque zone of obsidian and hot springs, otherwise known as La Primavera. A large 11 km diameter caldera was formed about 95,000 years ago, with the Tala Tuff being emplaced which has a volume of about 20 km^3 [31]. The caldera was then filled by a lake. Subsequent activity included the formation of many domes, which along with uplift ended the life of the lake. Its legacy can be seen in the form of a notorious deposit of so-called giant pumices, some of which are more than 1 m across [32]. The exploitation of the remnant heat for geothermal energy was commenced but never made it to production due to objections due to the environmental impact.

Volcán de Colima

Being the most active volcano in North America, Volcán de Colima deserves special attention. It currently is undergoing its most active period since the last major eruption, which occurred in 1913. The Colima Volcanic Complex shows a southern migration of activity, the oldest edifice being Cántaro which was active between 1.6 and 1 million years before present. Volcanism then moved 15 km south to Nevado de Colima, and more recently to the current location, a further 6 km to the south [33]. The existence of a group of undated domes, called Los Hijos about 3 km further south, possibly corresponds to the first signs of further migration. Some major debris avalanches have occurred from the edifice, including a large one 18,500 years ago, which formed a dam across the Naranjo river [34]. The breaking of this dam produced an enormous debris flow which reached the Pacific Ocean, 120 km away. More recently,

3,600 years ago, a smaller collapse went toward the SW, this time damming the Armería river [35]. The relatively large frequency of such events is demonstrated by evidence of a more recent event still, which has been dated at 2,500 years ago. Emergency response plans cannot consider events of such large magnitude. The probability of occurrence remains low; however, sooner or later a debris avalanche will occur, if not at Colima, at one of Mexico's other sizable stratovolcanoes.

Since the most recent collapse, pyroclastic deposits and lava flows have been accumulating and building the current edifice. It has grown to a height of about 3,860 m, though the current altitude is not really known due to the frequent construction and destruction of summit domes. The 1913 Plinian eruption produced a 23 km altitude column and pyroclastic flows that reached 15 km from the volcano [36]. The historical records show that Volcán de Colima has a large Plinian (or at least sub-Plinian) eruption every 100 years, more or less [37]. Previous to 1913 there were eruptions in 1818, 1690, and 1606. This is very frequent for a volcano.

Volcán de Colima has produced a large number of andesitic domes and flows during recent years. The most recent episode of activity commenced in 1998 [38] and has since produced four different phases of dome growth (1998–1999, 2001–2003, 2004, 2007–2011 [effusive eruption still ongoing at the time of writing: June 2011]). The effusive activity has been interspersed with explosive eruptions (Fig. 5). This peaked in 2005, when at least 30 Vulcanian eruptions occurred, each producing a pyroclastic flow resulting from column collapse. The longest flow reached 5.4 km along a ravine to the SE of the volcano. An even longer flow occurred in October 2004 when a relatively large dome collapse produced a pyroclastic flow which headed down a ravine on the SW flank to a point 6.1 km from the volcano. Since 2003, the volcano has produced several Vulcanian explosions each day (like the example in Fig. 6).

The population occupying the flanks of this volcano is much larger today than it was at the time of the last large eruption. Monitoring efforts include a variety of techniques applied in the quest to identify precursors of any large-scale acceleration of activity.

Volcanoes of Mexico. Figure 5

Volcán de Colima in January 2002. In the photo the older part of the complex, Nevado de Colima can be seen behind to the north. The volcano itself is covered with lava flows of different ages, the longest recent flows to the south having formed in 1998–1999. The black-looking lava dome had been growing since May 2001 and reached the crater rim and began forming rockfalls the following month. The Los Hijos domes can be seen in the bottom right-hand corner of the photo

Volcanoes of Mexico. Figure 6
Small typical Vulcanian explosion at Volcán de Colima, February 2009

The seismic network had its first seismometers installed in the 1980s. Useful precursors have been identified like swarms of long-period events prior to Vulcanian explosions or effusion (2004 and 2005; [39]) and volcanotectonic events signaling the ascent of magma in 1997–1998. The recent introduction of monitoring methods such as the thermal monitoring of fumaroles [40], the dome and explosions have expanded the possibilities of forming models to explain the transition between different regimes of activity, which form various cycles of activity: from the daily explosive cycle, the several year cycle of effusive episodes, to the 100 year cycle of large explosions. The integration of monitoring data will be vital to try and determine whether a cataclysmic eruption is imminent.

Central Stratovolcanoes

The Eastern-central zone of the TMVB features five large stratovolcanoes, four of which are considered active, along with a number of smaller edifices like Jocotitlán, also considered active. Here the descriptions start to the west and head east.

Nevado de Toluca

Nevado de Toluca reaches 4,680 m above sea level and features a pair of lakes within its caldera. Fig. 7 shows the smaller Laguna de la Luna (moon) within the summit caldera. The lakes have proved to be interesting to archaeologists, whose finds suggest they were used for prehispanic rituals. Its final episode of effusive activity 9,100 years ago produced a small dacitic lava dome. This followed the most recent major Plinian eruption, which produced the Upper Toluca Pumice deposit 10,500 years ago [41]. This major eruption produced a column that is thought to have reached 42 km in altitude and deposited about 14 km^3 of material. This was the largest eruption in Mexico in the last 15,000 years. An extensive record of large explosive activity and major collapses of the edifice can be observed in the surrounding region. The most recent deposits date about 3,300 years before present and consist of a pyroclastic flow and surge emplaced on the NE flank.

Jocotitlán

Nearby the smaller Jocotitlán is a fine example of an edifice that has collapsed to produce a debris avalanche deposit. In contrast to the larger stratovolcanoes, it rises only 1,300 m above the surrounding plains. Large conical hummocks have been identified with the collapse event that occurred 9,690 years ago [42].

Volcanoes of Mexico. Figure 7
Laguna de la Luna, within the crater of Nevado de Toluca

The 2.8 km^3 deposit reached a distance of 12 km. The most recent eruption of Jocotitlán was only 680 years ago; the material deposited from pyroclastic density currents on the upper flanks of the edifice makes it obligatory to include this peak in the list of active volcano of Mexico.

Popocatépetl

The potentially biggest volcanic threat in Mexico is Popocatépetl, which at 5,472 m towers above the surroundings, and is only 65 km from Mexico City and 45 km from Puebla (Fig. 8). The population within a radius of 40 km is in excess of one million. Popocatépetl reawoke in December of 1994. Since then activity has been characterized by the growth of lava domes within the crater, periodic Vulcanian explosions and the release of large volumes of gas. This eruption has been the first since El Chichón in 1982 to produce fatalities. Unfortunately five intrepid individuals decided to ignore official warnings and they climbed to the crater rim in April 1996. As well as ending their lives, an explosion sent cm-sized fragments of the destroyed lava dome to the nearest villages.

During the last 10,000 years there have been at least three large Plinian eruptions [43]. 4,965 years ago an eruption sent clasts with a diameter of up to 2.5 cm to a distance of 19 km from the volcano [44]. This implies that the height of the eruption column was 37–41 km. The eruption deposited 4.9 km^3 of material and was the second largest in Mexico within the Late Pleistocene–Holocene period. A similar eruption today would have an enormous impact on some 15 million people living in the surrounding region. More recently, two further large eruptions occurred 2,150 and 1,100 years ago. These two would have affected the local population centres at this time.

During the current activity there have been two reasonably sized lahar events in 1997 and 2001 [45]. The first had more water with a partial melting of the glacier making a contribution. Both events reached the most vulnerable village on the flanks: Santiago Xalitzintla, some 15 km from the crater, though nobody was hurt.

La Malinche

The indigenous wife of Cortés certainly would not have guessed that her name, or the Mexican nickname for her, would one day be given to a large stratovolcano. Like many volcanoes in this country, activity is punctuated by large periods of rest [46]. La Malinche is one

Volcanoes of Mexico. Figure 8
Popocatépetl, which gets a frequent covering of snow during the rainy season. Its permanent glacier has greatly diminished in size over recent years

volcano that has not appeared on lists of active Mexican volcanoes, explained by its lack of a clear crater or fumaroles. However, the local land is covered by pyroclastic deposits with poor soil development on top, and one layer has been dated at 3,100 years old. Given that over two million people live on its lower slopes, it needs to be considered carefully, with eruption scenarios determined [46].

Citlaltépetl

Citlaltépetl has the honor of being the highest active volcano of North America. Its prehispanic name is perhaps less used than the name which resulted from the Spanish invasion: Pico de Orizaba. At 5,675 m it is no mean feat to cross the glaciers and reach the rim of the crater. In recent years it has not shown much activity; however, over 750,000 people live within a radius of 40 km. The last major eruption, which occurred 4,100 years ago, produced a series of block and ash flows and lahars whose deposits have been found up to 28 km from the crater [47]. The most recent event was a smaller eruption that left deposits of tephra, having been dated at 690 years ago [48].

Like many of Mexico's volcanoes, widespread devastation is one possible future scenario in the event of a debris avalanche. There is evidence that Citlaltépetl has suffered many such events possibly without any eruptive activity [49], with some reaching the Gulf of Mexico, 120 km away. The edifice of this volcano has often been weakened by intense hydrothermal alteration, which combined with a large elevation difference of 4,400 m dropping down to the Gulf Coastal Plain have produced favourable conditions for these large-scale edifice collapses. This lack of precursory eruptive activity means that an event could occur one day without any warning.

Las Cumbres

Las Cumbres, located only about 10 km north of Citlaltépetl volcano, is an eroded stratovolcano which was once possibly as large as its neighbour [50]. It is the middle member of the enormous N-S chain of volcanoes which starts with Cofre de Perote in the north and ends with Citlaltépetl to the south. Its last big eruption was 20,000 years ago and produced the widespread Quetzalapa pumice deposit. More recently various scoria cones were formed as well as the Yolotepec dome,

Volcanoes of Mexico. Figure 9
The church of San Juan Parangaricutiro which was buried by the lava flows from Paricutín. The cone can be seen in the background

which has been dated at less than 6,000 years [50]. Large debris avalanches have also resulted from major collapses within this complex [49].

Volcanic Fields

Michoacán-Guanajuato Volcanic Field

Driving through this large expanse of territory, one cannot escape the awe-inspiring impact of the extent of this volcanic field. In places the cinder cones are so close they are touching one another. The recent addition of Paricutín is the most famous, but previously the birth of Jorullo was also witnessed in 1759. The field covers a vast area measuring 250 × 200 km in the two states which give the field its name. It contains at least 1,040 volcanic vents. The majority are cinder cones, but there are also small shield volcanoes, lava domes and maars. In general, the cones are randomly distributed, although there are local areas where alignments can be identified [51]. The region with the largest density of cones is that of Paricutín, where the median spacing between each one is 1.15 km. The cones are not so large, the median height being 90 m with a basal diameter of 800 m. A visit to Paricutín today awards the visitor with views of the famous church which was partly spared by the advancing lava flows. Figure 9 shows the church partially submerged.

One northern region of the field is Valle de Santiago, which features the stunning beauty of its seven major maars. There are a total of 20 within a zone with dimensions of 7 × 50 km [52]. The youngest, La Alberca, has been dated at 73,000 years old. Unfortunately, overexploitation of the groundwater has meant that the maars have almost all lost their lakes over recent years.

Durango Volcanic Field

This large field covers some 2,100 km^2 and contains about 100 cinder cones [53]. The La Brefia-El Jagüey Maar Complex is one of the youngest centres in the field and the only section studied in detail. Being maars, they were formed by the interaction between magma and groundwater (phreatomagmatic eruption), which produced large explosions. The final phase of the activity was the formation of several scoria cones within the crater and related lava flows. The age of this complex has been estimated at a few thousand years [53].

Mascota Volcanic Field

Several volcanic fields are located within the Jalisco Block, the youngest being close to the city of Mascota. This field is notorious for the geochemistry of its lavas, which are dominated by minettes, an unusual type of lava which, instead of the more common feldspar crystals dominating, contains large mica crystals [54]. The youngest flow of this field stands out through the lack of vegetation or soil on its surface. This led the authors to believe its age to be a few thousand years old. Further evidence of its age could be a correlation with scoria found in nearby lake deposits that was dated at less than 5,600 years old.

Chichinauzin Volcanic Field

The extensive Chichinauzin field lies to the south of Mexico City and covers some 2,500 km^2. It contains more than 200 monogenetic scoria cones and associated lava flows [55]. The last eruption was 1,670 years ago [56] and given that its historic <1,250 years reoccurrence time has been greatly exceeded, it represents one of the more likely regions for the birth of the next Mexican volcano. During the past 10,000 years, at least six monogenetic eruptions have occurred (Pelado, Cuauhtzin, Tláloc, Guespalapa, Chichinautzin, and Xitle). The cone of Xitle is a clear landmark within the southern limits of the city boundary, reminding us of its eruption 1,670 years ago. Obviously when the day arrives for the next eruption, the effect will be catastrophic and resulting lava flows and ash fall will paralyze the capital city.

Serdán-Oriental Volcanic Field

No evidence has been found for activity within the last 10,000 years; however, it has been included in Table 1 for two main reasons. Firstly, Las Derrumbadas are a pair of volcanic domes within this field. They still have active fumaroles outputting measurable quantities of sulphur dioxide. This volcanic field also hosts several maars, some with lakes, some without, other rhyolitic domes and scoria cones. An interesting dome named Cerro Pizarro shows evidence that it has produced multiple eruptions, unusual for this type of volcano, which normally is monogenetic [57]. Furthermore, the repose periods could be greater than 65,000 years, which is another reason for not declaring this field as no longer capable of producing an eruption.

Cofre de Perote Vent Cluster and Naolinco Volcanic Field

These recently identified mafic fields are located on the flank of the largely Pleistocene Cofre de Perote shield volcano, in the case of the Cofre de Perote Vent Cluster (CPVC) and to the north of Jalapa, Veracruz, the Naolinco Volcanic Field (NVF) [58]. The CPVC consists of an extensive lava field that covers >100 km^2. The most recent eruption was from the El Volcancillo scoria cone and it produced an impressively large flow which traveled 50 km only about 870 years ago. The Rincón de Chapultepec scoria cone in the NVF produced a lava flow ~2,980 BP. Interestingly these flows are some of the largest in the whole TMVB, but remained almost totally ignored until within the last 10 years.

San Martín Tuxtla

The basaltic volcano San Martín Tuxtla is located near the coast of the Gulf of Mexico, in southern Veracruz. It actually represents the active centre within another volcanic field containing monogenetic volcanic cones, maars, and three other large volcanoes which have not shown evidence of activity in the Holcene [59]. It is so far unclear whether this volcano is the eastward end-member of the TMVB or related to extensional tectonics. Its most recent eruption in 1793 was a succession from phreatomagmatic explosions, to Strombolian explosions and then an effusive episode, which produced a 3 km lava flow. At least nine other eruptions took place within the last 6,000 years of scoria cones and maars in this field.

Volcanoes of Chiapas

El Chichón

The largest historic eruption to occur in Mexico was the 1982 eruption of El Chichón, which resulted in the death of around 2,000 people and impacted the world's climate through a measurable reduction in temperature. Prior to the eruption it was almost totally unknown, with nobody aware of its potential for

devastation. El Chichón is not a towering stratovolcano like those of the TMVB, but an unassuming hill, which is difficult to see until one has almost arrived at its lower flanks. Ironically, several months before the eruption, El Chichón was the subject of a geothermal prospecting trip. In a report it was stated that the volcano needed to be studied and might reactivate in the future.

Three large explosions occurred within a week from 28 March to 4 April, 1982. The authorities were taken by surprise [60]. The volcano had been asleep for 550 years previous to this Plinian eruption. Lack of experience largely affected the handling of the emergency, with a proportion of the deaths occurring due to the early return of evacuees. The eruption attracted a lot of international interest, one reason being that the magma had an unusually high sulphur content. This increased the impact on the world's climate from the eruption cloud, with aerosols in the stratosphere producing a decrease of 0.2–0.5°C in temperature in the Northern Hemisphere [61].

The eruption at El Chichón formed a 1 km diameter and 180 m deep crater (Fig. 10). Geochemical and geophysical studies have attempted to define the hydrothermal system which includes the shallow crater lake, bubbling springs, and geysers [62, 63]. The lake can be observed to vary considerably in size, seemingly following a cycle that is independent of the local rainfall. Another interesting observation is that some hot springs found outside of the crater appear to be part of a system that was undisturbed by the Plinian eruption.

Research has now shown that the volcano has had a large eruption approximately every 800 years [64]. Studies looking at the stratigraphy of local deposits have deduced that during the past 8,000 years it has produced at least 11 eruptions [65], with many being similar to that of 1982. Larger magnitude events occurred in the year 750 and 1450 AD. The former of the two coincided with the collapse of the Mayan civilization in that region. Perhaps it played a role. A future large eruption would impact more than 70,000 people living in a radius of 35 km from the volcano.

Tacaná

This volcano sits right on the Mexico–Guatemala border and presents a largely unacknowledged hazard. The most recent eruption was in 1986: the phreatic explosion resulted in the formation of a new fumarole field

Volcanoes of Mexico. Figure 10
The crater of El Chichón taken on 5 May 2004

below the summit. Although it was a small event, it indicates the proximity of a magma body to the surface. The study of the geochemical characteristics of a number of springs at different elevations has provided an interesting insight into the often little-understood interaction between magmatic gases and the aquifers within the volcanic edifice [66].

The new fumaroles are located in a scar that was left by a collapse that occurred about 10,610 years ago after the growth of a summit dome [67]. More recently during the Holocene Tacaná had a series of eruptions, both explosive and effusive, the most recent being about 1,950 years ago [68]. Following the eruption, a series of lahars devastated the surrounding countryside. There is evidence that the construction of a nearby prehispanic settlement called Izapa was interrupted by these events. New activity of a similar magnitude could impact the approximately 300,000 people who live within 35 km of the volcano, an order of magnitude higher than its more famous neighbour in Chiapas.

Future Directions

Hopefully this entry has provided a useful introduction to the many volcanoes of Mexico. The cultural and geographical richness of the country is exemplified in its volcanoes. For their study, Mexico offers a huge variety of landforms and endless geophysical and geochemical case studies. There is a considerable potential for future eruptions with a significant impact. The most dangerous volcanoes in Mexico for their potential for large eruptions are probably Popocatépetl, Volcán de Colima, Tacaná, Citlaltépetl, and Ceboruco. But monogenetic fields such as Chichinautzin could also wreak havoc with the birth of a new cinder cone. The reactivation of one of its other sleeping giants always remains a distinct possibility.

Studies are only just being initiated of many of the active or potentially active volcanic centres of Mexico. There are many areas of research to be explored, with many more waiting to be defined. In certain cases the origin of extensive deposits emplaced during historic or prehistoric eruptions is unknown. With others the eruption mechanism has yet to be fully understood, which creates a dilemma when attempting to define a monitoring strategy, whilst faced with the possibility of an impending cataclysmic event. This lack of understanding makes it difficult to set threshold levels for monitored parameters, such as gas flux or seismicity, which should trigger a change in the alert system or an action within the emergency risk mitigation plan.

Hazard maps have been published of Popocatépetl, Volcán de Colima, Citlaltépetl, El Chichón, and Nevado de Toluca. These represent vital tools for the mitigation of volcanic risk by improving land-use planning and for defining procedures during emergencies. These need to be regarded as dynamic, with constant updating as new information becomes available. Work is underway to construct risk maps for certain hazards, which not only consider the geographical extent, but also the probability of occurrence and the vulnerability of the local population. Similar maps are required for some of the other active volcanoes of Mexico.

After the large loss of life at the hands of El Chichón, much progress has been made to reduce the risk at many of Mexico's volcanoes. More needs to be done in areas such as education to prepare for future potentially devastating eruptions.

Bibliography

Primary Literature

1. Capra L, Macías JL, Scott KM, Abrams M, Garduño-Monroy VH (2002) Debris avalanches and debris flows transformed from collapses in the Trans-Mexican Volcanic Belt, Mexico – behavior, and implications for hazard assessment. J Volcanol Geotherm Res 113:81–110
2. Anguita F et al (2001) Circular features in the Trans-Mexican Volcanic Belt. J Volcanol Geotherm Res 107(4):265–274
3. Asociación Geotérmica Mexicana. http://www.geotermia.org.mx. Accessed June 2011
4. Medina F, Suarez F, Espindola JM (1989) Historic and holocene volcanic centers in NW Mexico. Bull Volcanol 51:91–93
5. Pardo M, Suárez G (1995) Shape of the subducted Rivera and Cocos plates in southern Mexico: seismic and tectonic implications. J Geophys Res 100(B7):12357–12373
6. Ferrari L (2004) Slab detachment control on mafic volcanic pulse and mantle heterogeneity in central Mexico. Geology 32(1):77–80
7. Márquez A, Oyarzun R, de Ignacio C, Doblas M (2001) Southward migration of volcanic activity in the central Mexican Volcanic Belt: asymmetric extension within a two-layer crustal stretching model. J Volcanol Geotherm Res 112(1–4):175–187
8. Manea M, Manea VC (2008) On the origin of El Chichón volcano and subduction of Tehuantepec Ridge: a geodynamical perspective. J Volcanol Geotherm Res 175(4):459–471

9. Luhr JF, Nelson SA, Allan JF, Carmichael ISE (1985) Active rifting in Southwestern Mexico: manifestations of an incipient eastward spreading-ridge jump. Geology 13:54–57
10. Frey HM, Lange RA, Hall CM, Delgado-Granados H, Carmichael ISE (2007) A Pliocene ignimbrite flare-up along the Tepic-Zacoalco rift: evidence for the initial stages of rifting between the Jalisco Block (Mexico) and North America. GSA Bull 119(1/2):49–64
11. Gutmann JT (2002) Strombolian and effusive activity as precursors to phreatomagmatism: eruptive sequence at maars of the Pinacate volcanic field, Sonora, Mexico. J Volcanol Geotherm Res 113(1–2):345–356
12. de Boer J (1980) Paleomagnetism of the Quaternary Cerro Prieto, Crater Elegante, and Salton Buttes volcanic domes in the northern part of the Gulf of California rhombochasm, 2. symposium on the Cerro Prieto Geothermal Field. Baja California, Mexico, p 10
13. Luhr JF (1995) San Quintin volcanic field, Baja California Norte, Mexico: geology, petrology, and geochemistry. J Geophys Res 100(B7):10353–10380
14. Ortega-Rivera A, Bohnel H, Lee J (2004) The San Quintin volcanic field – $^{40}Ar/^{39}Ar$ geochronology and paleomagnetism. Geological Society of America Penrose Conference, Metepec, Puebla, p 59
15. Rogers G, Saunders AD, Terrell DJ, Verma SP, Marriner GF (1985) Geochemistry of Holocene volcanic rocks associated with ridge subduction in Baja California, Mexico. Nature 315:389–392
16. Capra L, Macías JL, Espíndola JM, Siebe C (1998) Holocene plinian eruption of La Virgen volcano, Baja California, Mexico. J Volcanol Geotherm Res 80:239–266
17. Paz Moreno FA, Demant A (1999) The recent Isla San Luis volcanic centre: petrology of a rift-related volcanic suite in the northern Gulf of California, Mexico. J Volcanol Geotherm Res 93(1–2):31–52
18. Batiza R (1978) Geology, petrology, and geochemistry of Isla Tortuga, a recently formed tholeiitic island in the Gulf of California. Geol Soc Amer Bull 89:1309–1324
19. Housh TB, Aranda-Gómez JJ, Luhr JF (2010) Isla Isabel (Nayarit, México): quaternary alkalic basalts with mantle xenoliths erupted in the mouth of the Gulf of California. J Volcanol Geotherm Res 197(1–4):85–107
20. Bohrson WA et al (1996) Prolonged history of silicic peralkaline volcanism in the eastern Pacific Ocean. J Geophys Res 101(B5):11457–11474
21. Richards AF (1959) Geology of the Islas Revillagigedo, Mexico 1. Birth and development of Volcan Barcena, Isla San Benedicto. Bull Volcanol 22:73–124
22. Siebe C et al (1995) Submarine eruption near Socorro Island, Mexico: geochemistry and scanning electron microscopy studies of floating scoria and reticulite. J Volcanol Geotherm Res 68:239–271
23. Taran YA, Varley NR, Inguaggiato S, Cienfuegos E (2010) Geochemistry of H_2- and CH_4-enriched hydrothermal fluids of Socorro Island, Revillagigedo Archipelago, Mexico. Evidence for serpentinization and abiogenic methane. Geofluids 10:542–555
24. Farmer JD, Farmer MC, Berger R (1993) Radiocarbon ages of lacustrine deposits in volcanic sequences of the Lomas Coloradas area, Socorro Island, Mexico. Radiocarbon 35(2):253–262
25. Ferrari L, Pasquarè G, Venegas-Salgado S, Romero-Rios F (1999) Geology of the western Mexican volcanic belt and adjacent Sierra Madre Occidental and Jalisco Block. Cenozoic tectonics and volcanism of Mexico. Geological Society of America, Special paper 334, pp 65–83
26. Gardner JE, Tait S (2000) The caldera-forming eruption of Volcán Ceboruco, Mexico. Bull Volcanol 62:20–33
27. Sieron K, Siebe C (2008) Revised stratigraphy and eruption rates of Ceboruco stratovolcano and surrounding monogenetic vents (Nayarit, Mexico) from historical documents and new radiocarbon dates. J Volcanol Geotherm Res 176(2): 241–264
28. Nelson SA, Carmichael ISE (1984) Pleistocene to recent alkalic volcanism in the region of Sanganguey volcano, Nayarit, Mexico. Contrib Mineral Petrol 85(4):321–335
29. Luhr JF (2000) The geology and petrology of Volcán San Juan (Nayarit, México) and the compositionally zoned Tepic Pumice. J Volcanol Geoth Res 95(1–4):109–156
30. Lewis-Kenedi CB, Lange RA, Hall CM, Delgado-Granados H (2005) The eruptive history of the Tequila volcanic field, western Mexico: ages, volumes, and relative proportions of lava types. Bull Volcanol 67(5):391–414
31. Mahood GA (1981) A summary of the geology and petrology of the Sierra La Primavera, Jalisco, Mexico. J Geophys Res 86(B11):10137–10152
32. Walker GPL, Wright JV, Clough BJ, Booth B (1981) Pyroclastic geology of the rhyolitic volcano of La Primavera, Mexico. Geol Rundschau 70:1100–1118
33. Luhr JF, Carmichael ISE (1980) The Colima volcanic complex, Mexico. 1. Post-caldera andesites from Volcan Colima. Contrib Minerol Petrol 71:343–372
34. Capra L, Macías JL (2002) The cohesive Naranjo debris-flow deposit (10 km^3): a dam breakout flow derived from the Pleistocene debris-avalanche deposit of Nevado de Colima Volcano (Mexico). J Volcanol Geotherm Res 117(1–2):213–235
35. Cortés A, Macías JL, Capra L, Garduño-Monroy VH (2010) Sector collapse of the SW flank of Volcán de Colima, México: the 3600 yr BP La Lumbre-Los Ganchos debris avalanche and associated debris flows. J Volcanol Geotherm Res 197(1–4): 52–66
36. Saucedo R et al (2010) Eyewitness, stratigraphy, chemistry, and eruptive dynamics of the 1913 Plinian eruption of Volcán de Colima, México. J Volcanol Geotherm Res 191(3–4):149–166
37. Luhr JF (2002) Petrology and geochemistry of the 1991 and 1998–1999 lava flows from Volcán de Colima, Mexico: implications for the end of the current eruptive cycle. J Volcanol Geotherm Res 117(1–2):169–194

38. Zobin VM et al (2002) Overview of the 1997–2000 activity of Volcán de Colima, Mexico. J Volcanol Geotherm Res 117(1–2): 1–19
39. Varley N, Arámbula-Mendoza R, Reyes-Dávila G, Stevenson J, Harwood R (2010) Long-period seismicity during magma movement at Volcán de Colima. Bull Volcanol 72(9): 1093–1107
40. Stevenson JA, Varley N (2008) Fumarole monitoring with a handheld infrared camera: Volcán de Colima, Mexico, 2006–2007. J Volcanol Geotherm Res 177(4):911–924
41. Arce JL, Macías JL, Vázquez-Selem L (2003) The 10.5 ka Plinian eruption of Nevado de Toluca volcano, Mexico: stratigraphy and hazard implications. GSA Bulletin 115(2):230–248
42. Siebe C, Komorowski J-C, Sheridan MF (1992) Morphology and emplacement of an unusual debris avalanche deposit at Jocotitlán volcano, central Mexico. Bull Volcanol 54:573–589
43. Siebe C, Abrams M, Macías JL, Obenholzner J (1996) Repeated volcanic disasters in Prehispanic time at Popocatépetl, central Mexico: past key to the future? Geology 24:399–402
44. Arana-Salinas L, Siebe C, Macías JL (2010) Dynamics of the ca. 4965 yr ^{14}C BP "Ochre Pumice" Plinian eruption of Popocatépetl volcano, México. J Volcanol Geotherm Res 192(3–4): 212–231
45. Capra L, Poblete MA, Alvarado R (2004) The 1997 and 2001 lahars of Popocatépetl volcano (Central Mexico): textural and sedimentological constraints on their origin and hazards. J Volcanol Geotherm Res 131(3–4):351–369
46. Castro-Govea R, Siebe C (2007) Late Pleistocene-Holocene stratigraphy and radiocarbon dating of La Malinche volcano, Central Mexico. J Volcanol Geotherm Res 162(1–2):20–42
47. Carrasco-Núñez G (1999) Holocene block-and-ash flows from summit dome activity of Citlaltepetl volcano, Eastern Mexico. J Volcanol Geotherm Res 88(1–2):47–66
48. De la Cruz-Reyna S, Carrasco-Núñez G (2002) Probabilistic hazard analysis of Citlaltépetl (Pico de Orizaba) Volcano, eastern Mexican Volcanic Belt. J Volcanol Geotherm Res 113(1–2):307–318
49. Carrasco-Núñez G et al (2006) Multiple edifice-collapse events in the Eastern Mexican Volcanic Belt: the role of sloping substrate and implications for hazard assessment. J Volcanol Geotherm Res 158:151–176
50. Rodriguez SR (2005) Geology of Las Cumbres Volcanic Complex, Puebla and Veracruz states, Mexico. Rev Mex Cienc Geol 22(2):181–199
51. Hasenaka T, Carmichael ISE (1985) The cinder cones of Michoacán-Guanajuato, Central Mexico: their age, volume and distribution, and magma discharge rate. J Volcanol Geotherm Res 25:105–124
52. Uribe-Cifuentes RM, Urrutia-Fucugauchi J (1999) Paleomagnetic study of the Valle de Santiago volcanics, Michoacán-Guanajuato volcanic field, Mexico. Geofís Intern 38(4):217–230
53. Aranda-Gómez JJ, Luhr JF, Pier G (1992) The La Breña – El Jaguey Maar Complex, Durango, México: I. Geological evolution. Bull Volcanol 54(5):393–404
54. Carmichael ISE, Lange RA, Luhr JF (1996) Quaternary minettes and associated volcanic rocks of Mascota, western Mexico: a consequence of plate extension above a subduction modified mantle wedge. Contrib Mineral Petrol 124(3):302–333
55. Siebe C, Rodríguez-Lara V, Schaaf P, Abrams M (2004) Radiocarbon ages of Holocene Pelado, Guespalapa, and Chichinautzin scoria cones, south of Mexico city: implications for archaeology and future hazards. Bull Volcanol 66:203–225
56. Siebe C, Arana-Salinas L, Abrams M (2005) Geology and radiocarbon ages of Tlaloc, Tlacotenco, Cuauhtzin, Hijo del Cuauhtzin, Teuhtli, and Ocusacayo monogenetic volcanoes in the central part of the Sierra Chichinautzin, Mexico. J Volcanol Geotherm Res 141(3–4):225–243
57. Carrasco-Núñez G, Riggs NR (2008) Polygenetic nature of a rhyolitic dome and implications for hazard assessment: Cerro Pizarro volcano, Mexico. J Volcanol Geotherm Res 171(3–4):307–315
58. Siebert L, Carrasco-Núñez G (2002) Late-Pleistocene to precolumbian behind-the-arc mafic volcanism in the eastern Mexican volcanic belt; implications for future hazards. J Volcanol Geotherm Res 115(1–2):179–205
59. Espíndola JM, Zamora-Camacho A, Godinez ML, Schaaf P, Rodríguez SR (2010) The 1793 eruption of San Martín Tuxtla volcano, Veracruz, Mexico. J Volcanol Geotherm Res 197(1–4): 188–208
60. Tilling RI (2009) El Chichón's "surprise" eruption in 1982: lessons for reducing volcano risk. Geofís Intern 48(1):3–19
61. Luhr JF, Carmichael ISE, Varekamp JC (1984) The 1982 eruptions of El Chichón Volcano, Chiapas, Mexico: mineralogy and petrology of the anhydritebearing pumices. J Volcanol Geotherm Res 23(1–2):69–108
62. Rouwet D, Taran Y, Inguaggiato S, Varley N, Santiago Santiago JA (2008) Hydrochemical dynamics of the "lake-spring" system in the crater of El Chichón volcano (Chiapas, Mexico). J Volcanol Geotherm Res 178(2):237–248
63. Jutzeler M, Varley N, Roach M (2011) Geophysical characterization of hydrothermal systems and intrusive bodies, El Chichón volcano (Mexico). J Geophys Res 116(B4):B04104
64. Macías JL et al (2008) Hazard map of El Chichón volcano, Chiapas, México: constraints posed by eruptive history and computer simulations. J Volcanol Geotherm Res 175(4):444–458
65. Espíndola JM, Macías JL, Tilling RI, Sheridan MF (2000) Volcanic history of El Chichón volcano (Chiapas, Mexico) during the Holocene, and its impact on human activity. Bull Volcanol 62(2):90–104
66. Rouwet D, Inguaggiato S, Taran Y, Varley N, Santiago SJ (2009) Chemical and isotopic compositions of thermal springs, fumaroles and bubbling gases at Tacaná volcano (Mexico-Guatemala): implications for volcanic surveillance. Bull Volcanol 71(3):319–335
67. Macías J et al (2010) Late-Pleistocene flank collapse triggered by dome growth at Tacaná volcano, México-Guatemala, and its relationship to the regional stress regime. Bull Volcanol 72(1):33–53
68. Macías JL et al (2000) Late Holocene Peléan style eruption at Tacaná volcano, Mexico-Guatemala: past, present, and future hazards. Geol Soc Am Bull 112:1234–1249

Books and Reviews

Delgado-Granados H, Aguirre-Díaz G, Stock JM (eds) (1999) Cenozoic tectonics and volcanism of Mexico. GSA special paper 334

Francis P, Oppenheimer C (2003) Volcanoes. Oxford University Press, Oxford, p 536

Houghton B, Rymer H, Stix J, McNutt S, Sigurdsson H (1999) Encyclopedia of volcanoes. Academic, San Diego, p 1417

Luhr JF, Varekamp JC (eds) (1984) El Chichón volcano, Chiapas, Mexico: special issue. J Volcanol Geotherm Res 23(1/2):1–191

Macías JL (2007) Geology and eruptive history of some active volcanoes of Mexico. Geol Soc Am Bull 422:183–232

Nelson SA (1990) Volcanic hazards in Mexico – a summary. Univ Nacional Autón México Inst Geol Rev 9(1):71–81

Parfitt L, Wilson L (2008) Fundamentals of physical volcanology. Wiley-Blackwell, Malden, p 256

Schminke H-U (2005) Volcanism. Springer, Berlin/Heidelberg, p 334

Volcanoes, Observations and Impact

Clifford Thurber[1], Stephanie Prejean[2]
[1]Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA
[2]Seismology, USGS Volcano Science Center, Alaska Volcano Observatory, Anchorage, AK, USA

Article Outline

Glossary

Caldera Large crater formed by collapse of an overlying structure when an eruption empties a magma reservoir.

Effusive Eruption of fluid molten material that later solidifies.

Fumarole A volcanic vent that emits hot gas.

Infrasound Sound waves at frequencies below the range of human hearing (<20 Hz).

Interferogram A pattern of satellite radar wave "fringes" formed by interference, analogous to the colorful pattern from light reflected by a thin film of oil or gas, that can indicate ground deformation.

Lahar Heavy flow of mud, water, and debris triggered by interactions of hot material with ice of water or when heavy rain falls on recently erupted unconsolidated material.

Phreatic Explosion caused by heating and expansion of ground water.

Pyroclastic Composed of rock fragments ejected explosively from an erupting volcano.

Tremor Continuous vibration of the ground related to magma movement.

Volatiles Dissolved gases contained in magma.

Definition of the Subject

Volcanoes are critical geologic hazards that challenge our ability to make long-term forecasts of their eruptive behaviors. They also have direct and indirect impacts on human lives and society. As is the case with many geologic phenomena, the time scales over which volcanoes evolve greatly exceed that of a human lifetime. On the other hand, the time scale over which a volcano can move from inactivity to eruption can be rather short: months, weeks, days, and even hours. Thus, scientific study and monitoring of volcanoes is essential to mitigate risk. There are thousands of volcanoes on Earth, and it is impractical to study and implement ground-based monitoring at them all. Fortunately, there are other effective means for volcano monitoring, including increasing capabilities for satellite-based technologies.

In recent history, the destructive power of earthquakes and earthquake-induced tsunamis has been quite salient. Over the centuries and millennia, however, volcanic eruptions and eruption-induced tsunamis have had impacts that rival those of earthquakes, and in some cases have had a global reach. A prime example is the 1815 eruption of Tambora in Indonesia. That eruption is blamed for the catastrophic "Year without a Summer," when global cooling due to reflection of the Sun's energy by aerosols and ash injected into the atmosphere during the eruption led to massive crop failures and many deaths from starvation around the world [1, 2]. The Earth has also not witnessed a massive caldera-forming

eruption, such as those that gave rise to Yellowstone and Long Valley calderas, since the formation of Toba caldera (also in Indonesia) about 75,000 years ago [3].

There is also a contrast between earthquakes and volcanoes in terms of predictability. Although reliable and effective earthquake prediction remains an elusive goal [4] and warning systems are operational in very few places (e.g., Japan), there are numerous examples of successful eruption forecasts and warnings. An example is the impressive success of the US Geological Survey's Volcano Disaster Assistance Program (VDAP), which reports dozens of successful eruption forecasts and warnings in the 25 years of the program's history [5]. Somewhat ironically, it is the monitoring of seismic activity that has been the key to VDAP's success.

Introduction

Although most of the world's magmatic activity occurs underneath the oceans, primarily along mid-ocean ridges, the discussion is restricted to volcanism on continents and islands. Within this subaerial class of volcanism, there are three main categories in terms of tectonic setting: subduction zones, hotspots, and continental rifts, with examples provided by Indonesia, Hawaii, and East Africa, respectively. Iceland is a special example of an above sea level section of mid-ocean ridge. The first-order classification of erupted products is based on silica content, with further distinctions based on alkali content (sodium and potassium) [6] and the size of crystallized mineral grains. More silica-rich (felsic) lavas are predominant at subduction zones and more silica-poor (mafic) lavas are predominant at hotspots and continental rifts. In general, mafic lavas erupt more effusively whereas felsic lavas are more prone to explosiveness. For a thorough introduction to types of volcanoes and lavas and their potential for explosivity, the interested reader is referred to Lockwood and Hazlett [7].

The magnitude and violence of volcanic eruptions can be quantified in several ways. A common measure that is used in this entry is the Volcanic Explosivity Index (VEI) [8]. Volume of erupted material, ash cloud height, eruption duration, and qualitative observations describing eruption intensity are used in the calculation of VEI. Worldwide eruptions to date have been classified as VEI 0-8, with each increasing integer corresponding to an order of magnitude increase in eruption severity. The 1980 and 2004–2008 eruptions of Mount St. Helens, for example, are classified as VEI 4 (large eruption) and VEI 2 (moderate eruption), respectively. The largest eruptions of the twentieth century were of VEI 6, including the 1912 eruption of Novarupta on the Alaska Peninsula and the 1991 eruption of Mt. Pinatubo, Philippines. A second commonly used measure of eruption size is the dense-rock equivalent (DRE) of erupted material. This parameter specifies the actual amount of magma erupted and is dependent on careful field studies of erupted deposits. The Smithsonian Museum of Natural History keeps an updated database of these two eruption size parameters for recent and historical eruptions (http://www.volcano.si.edu/world/). Pyle [9] summarizes these and other measures of eruption sizes.

In this entry, the focus is mainly on geophysical observations of volcanoes and calderas as they pertain to eruption forecasting and prediction. Some key aspects of the impacts that eruptions have on humans and selected aspects related to sustainability have also been characterized.

Observations

The Role of Geologic Mapping

Our focus is primarily on geophysical volcano monitoring, but the importance of geologic mapping and associated studies cannot be overstated in providing the background information necessary to interpret these data correctly. In the case of erupting volcanoes, the past is generally the key to the present. Volcanoes often erupt similar magmas in similar volumes; thus, geological mapping to determine eruption histories provides a framework in which to interpret renewed unrest at a previously quiet volcano. In the case when a volcano's behavior diverges from its historical activity, knowledge of eruptive history allows us to understand how the magma system is evolving with time. Geologic mapping also characterizes the spatial distribution of hazards from previous eruptions including tephra fall, lahars and pyroclastic and lava flows. These maps can have an important role in land use planning. In addition, careful petrologic, petrographic, and isotopic analyses of erupted material can provide valuable evidence regarding magma storage and transport

history. These analyses complement the geophysical measurements described below to characterize the magmatic system at depth. For further details, the interested reader is referred to Decker [10], Simkin and Siebert [11], and Lockwood and Hazlett [7].

Seismology

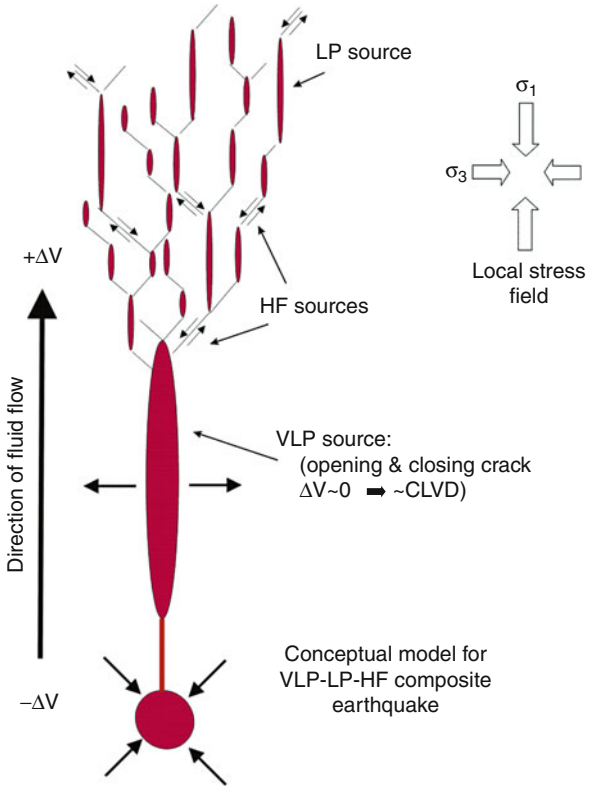
Earthquake monitoring is certainly one of the most basic and widely used techniques for observing volcanic activity [12, 13]. In fact, seismic monitoring of volcanoes has generally been the most fruitful approach for short-term eruption prediction. The successes of the US Geological Survey's Volcano Disaster Assistance Program (VDAP) [5] in using simple measures of earthquake activity to predict eruptions is a remarkable testament to the value of real-time seismic monitoring for taking the pulse of a volcano in a state of unrest. VDAP has assisted with or directly provided more than 50 successful eruption forecasts and/or predictions for more than 30 volcanoes worldwide in the first 25 years of its history (1986–2011), using simple instrumentation and basic observations of seismicity and ground shaking [5].

Earthquakes associated with volcanoes are commonly grouped into four classes: volcano-tectonic (VT, predominant frequencies of 1–20 Hz), long-period (LP, predominant frequencies of 1–5 Hz), hybrid (VT event transitioning into an LP event), and very-long-period (VLP) [14, 15]. The former are generally believed to be normal brittle-failure earthquakes, although they commonly occur in swarms of similar-sized small events rather than a main shock-aftershock sequence style. This can be quantified by the log frequency versus magnitude (i.e., Gutenberg-Richter) relation (i.e., b-value), which is characterized by a linear trend with a slope around 1 for earthquakes on crustal faults but frequently has a higher slope, up to roughly two, for volcanic swarms.

The mechanism of LP's remains a subject of debate, centered mainly on the effect of fluids on the earthquake source versus the effects of wave propagation (path effects), and the connection between LP's and the phenomenon known as volcanic tremor is also controversial. Volcanic tremor is a more or less continuous signal lasting minutes to hours to days, with a comparable frequency content to LP's. Some researchers believe volcanic tremor is simply the superposition of repeated LP events or, in the case of harmonic tremor, VT events with the observed frequency content controlled by the earthquake repetition rate [16]. A somewhat more common hypothesis is that the tremor is due to resonant oscillations in a magma conduit or a nonlinear response to fluid flow through cracks [17, 18]. The initiation of volcanic tremor is clearly associated with the movement of magma and the potential for an eruption, and thus tremor is one of the important precursors for eruption prediction [13]. In addition, deep LP's (typically 20–50 km depth) have been observed prior to a number of eruptions, presumably reflecting magma movement at depth [19–21], so additional focus has been placed on identifying these events. The nature of hybrid events is also debated, again centered on source versus path effects [22].

With the increasing use of broadband seismometers (instruments with a wide frequency range) in volcano monitoring, VLP earthquakes have been identified in many places [15, 23–38]. Similar to LP events, the source of VLP's is generally attributed to fluid-rock interaction, specifically transport of magma through the shallow crust. Waite et al. [38] for example, found that at Mount St. Helens, the VLP source is best modeled as a combination of volumetric and single-force components, the former due to compression and expansion of a shallow, magma-filled sill, and a smaller component of expansion and compression of a dike, and the latter due to mass transport in the magma conduit. A cartoon suggesting possible interrelationships among VLP, LP, and VT earthquakes based on the Hill fracture mesh concept [39] is shown in Fig. 1.

Some seismic path measurements have unveiled time dependencies that have been associated with eruptions, but for the most part these techniques have been applied retroactively. One of the earliest such studies was by Foulger et al. [40], who found changes in the ratio of the velocity of P waves (primary, or compressional) to S waves (secondary, or shear) (i.e., Vp/Vs, equivalently Poisson's ratio) at Mammoth Mountain, California, that correlated spatially and temporally with increased tree kill due to CO_2 emission. They hypothesized that an increased presence of gas in fractures led to a reduction in Vp/Vs, which was imaged using seismic tomography. Such changes in Vp/Vs have



Volcanoes, Observations and Impact. Figure 1
Cartoon of the hypothetical relationship among VLP, LP, and HF (i.e., VT) earthquakes. In this model, the VLP source is upward flow of magma with volume change ΔV. Volatiles from the magma permeate the crust above the VLP event, triggering high frequency (HF) and LP earthquakes in a stress field with most and least compressive stress directions, σ_1 and σ_3 respectively, as shown (Modified from [39])

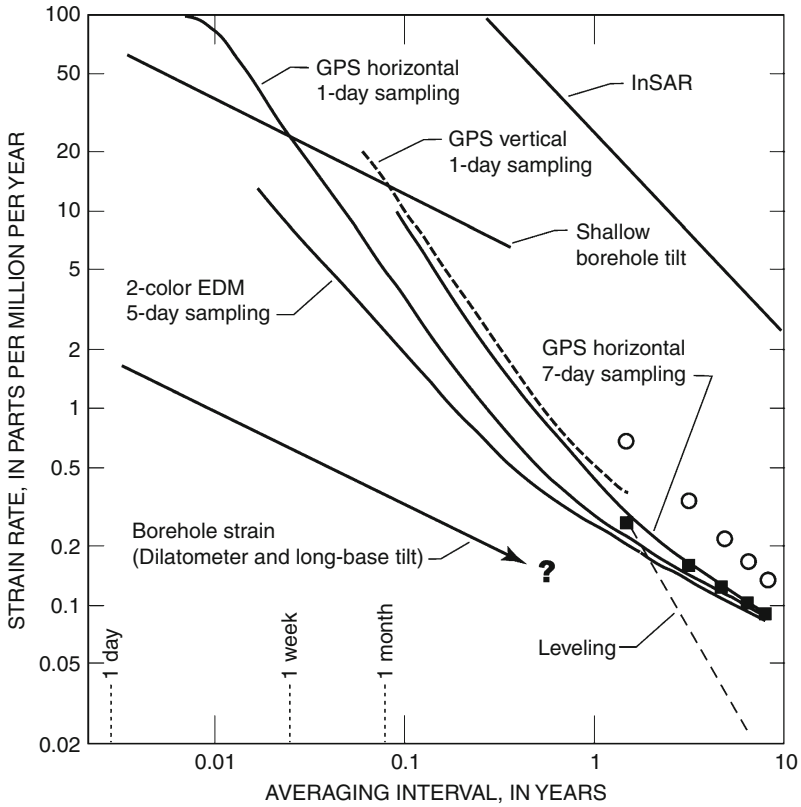
also been identified at Mt. Etna, Italy [41]. Recently, temporal changes in seismic wave attenuation, in this case associated with magmatic activity, have been found at Mt. Ruapehu, New Zealand [42] as well as Mt. Etna [43].

Two other types of seismic path observations that have been reported to show temporal change associated with magmatic activity are shear wave splitting, which is caused by birefringence in the Earth, and ambient noise correlations, which yield an estimate of the wave propagation behavior (the Green's function) between two seismic stations. Miller and Savage [44] identified a change in the polarization direction of the fast shear wave at Mt. Ruapehu, which they associated with an eruption. Subsequent studies have also reported such changes elsewhere related to magmatic activity [45–49], but extreme care must be taken to separate spatial versus temporal variations. Brenguier et al. [50] compared a reference Green's function (created from 18 months of background or "ambient" noise data) to Green's functions from sequential 10-day periods of data at Piton de la Fournaise Volcano on La Reunion island and found shifts in apparent travel times corresponding to a small reduction in seismic velocity within the volcano. Their interpretation was that decreases in seismic velocity a few weeks before eruptions were related to preeruptive inflation of the volcanic edifice, presumably due to increased magma pressure.

Deformation

Uplift and subsidence associated with magmatic and eruptive activity have been recognized for centuries, with the Temple of Serapis in the town of Pozzuoli in Campi Flegrei caldera, Italy, being one of the most famous examples [51]. Geodetic surveying techniques, such as leveling and tilt measurements, began to be applied to volcanoes in the early twentieth century [52]. In the modern era, Global Positioning System (GPS) and interferometric synthetic aperture radar (InSAR), along with tilt, are the primary types of geodetic observations applied to the study and monitoring of volcanoes. The sensitivity of these different techniques is compared in Fig. 2. Continuous GPS (cGPS) and InSAR are quite complementary, with the former providing fine temporal resolution at particular points on the Earth's surface and the latter providing exceptional spatial resolution of position changes over a wide area for pairs of observation times. Deformation modeling is commonly done to estimate pressure changes in and the geometry of magma source zones.

GPS Since the 1990s, GPS has become a central technique for volcano monitoring. Initial applications required repeated visits to benchmarks, known as campaign GPS. Continuous GPS (cGPS) has become relatively commonplace, allowing near-real-time tracking of site positions and, when multiple cGPS sites are available, a time history of volcano deformation.



Volcanoes, Observations and Impact. Figure 2

Comparison of rate sensitivity for GPS, InSAR, borehole strainmeters at 200 m depth, and borehole tiltmeters (which measure the gradient in vertical deformation) at 2 m depth. The x-axis indicates the time period that may be spanned by the different data types. For example, daily GPS measurements may span 1 day or longer. The y-axis indicates the strain rate that can be resolved as a function of the period. Strain is the change in length (area or volume) per a unit length (area or volume), and thus is unitless and can be expressed as parts per million (ppm). As can be seen from the plot, the borehole tiltmeters and strainmeters are more sensitive than GPS at shorter periods, but at periods longer than a few days and a few months, respectively, GPS measurements provide better resolution of strain rates (Figure modified from [52])

For example, with two cGPS sites on opposite sides of a volcano's summit or straddling a rift zone, relative changes in separation of the sites would reflect volcano inflation/deflation or rift zone extension/contraction. With enough sites, detailed modeling of deformation sources is possible, providing constraints on the locations, depths, and pressure/volume changes of buried magma reservoirs and conduits. An example of the use of cGPS in monitoring an impending eruption at Augustine Volcano, Alaska [53] is described in some detail later.

GPS works by having a receiver on the surface of the Earth receive signals simultaneously from four or more satellites. These signals carry information about the signal origin time and the position of the satellite, so that determining the receiver position is done by trilateration, essentially analogous to locating an earthquake. The interested reader is referred to [52] for a lucid and comprehensive discussion of the details of this process.

GPS is not without its complications and limitations. GPS is a point measurement of position, so good spatial coverage requires multiple GPS sites. Data reduction requires knowledge of the GPS satellite orbits, which are known only approximately at any instant. Final orbital information is distributed by

International GNSS Services (where GNSS stands for Global Navigation Satellite System) with about a 2-week delay [54]. The use of differential and kinematic GPS techniques effectively overcomes orbital (and some other) error effects, if only relative positions are desired. Atmospheric effects, due to ionosphere and troposphere delays, also impart positioning errors. Finally, elevation uncertainty is significantly greater than that for latitude and longitude.

InSAR InSAR burst onto the geophysics scene in 1993 with the publication of the famous Landers, CA, earthquake interferogram on the cover of Nature [55]. Applications to volcanoes and calderas around the world followed soon thereafter, including Mt. Etna, Italy [56], Long Valley caldera, CA [57], Yellowstone caldera, WY [58], Piton de la Fournaise Volcano, La Reunion [59], a number of volcanoes in Alaska [60–65], the Galapagos Islands [66], Afar [67], Chile [68], and many more.

InSAR works by having a satellite (or constellation of satellites) acquire a pair of radar images of the same area on Earth from about the same point in space at different times, which are then combined, or interfered, to produce a map of the difference in phase of the two returned radar signals, represented by colored "fringes." The phase difference is related to the line-of-sight (LOS; i.e., in the direction of the satellite) displacement of the ground in the time between the two image acquisitions. Converting the phase differences into a map of LOS surface displacement is a process known as "unwrapping" [69, 70]. As in the case of GPS, source models can be derived to fit the observed displacements.

The spatial extent and resolution of InSAR coverage is extraordinary (scenes typically 10s to 100 km on a side with 1–30 m pixel size and <cm resolution of surface displacements). There are many more subtleties and complexities with InSAR data and their interpretation than for GPS, however. Repeated image acquisitions are limited by the configuration and orbits of the satellites. An accurate digital elevation model (DEM) is critical for removing the effects of topography from the images. Noise in the data can lead to an incorrect assessment of LOS ground movement. Signal decorrelation, for example, due to vegetation variations, the presence of water, snow and ice, or steep topography, can yield areas with no usable signal. Atmospheric delay anomalies, due mainly to variations in tropospheric water content, can also cause artifacts in InSAR images. With the need for repeated imaging and the inherently limited temporal coverage for forming interferometric pairs, there is no guarantee that a good image can be created for a particular time interval of interest.

Tilt and Strain Although far less ubiquitous in volcano monitoring settings than seismic and GPS instrumentation, tiltmeters and strainmeters are valuable observational tools that can provide complementary information about deformation. Both have remarkable sensitivity, with standard (bubble) tiltmeters able to measure the equivalent of 0.1 mm of uplift over a range of a kilometer (0.1 μrad, or 10^{-7} strain). Strainmeters can measure strain on the order of 10^{-11} to 10^{-12} (10 to 1 parts per billion). Tiltmeters have a number of advantages over strainmeters, including much lower cost, simpler installation, and, for biaxial sensors, the ability to provide information on the direction to the source of deformation. These advantages have been responsible for the much greater use of tiltmeters than strainmeters by volcano observatories. Due to their excellent sensitivity to vertical deformation, in contrast to GPS with its more precise determination of horizontal position, and real-time capability, tiltmeters can be of great value in volcanic crisis situations [52]. Strainmeters have also proven their value in long-term observatory settings, though, such as the successful prediction of an eruption of Hekla Volcano, Iceland, in 2000 [71].

Gravity Gravity measurements can be used to infer vertical surface displacements with an accuracy comparable to GPS and some leveling techniques, but its real power comes from the ability to infer subsurface mass or density changes when surface deformation is constrained independently [52]. When uplift or subsidence occurs, the change in gravity can be compared to that predicted from the free-air gravity gradient. Deviations from the expected change in gravity reflect either an increase or decrease in mass in the subsurface [72], and the degree and sign of the difference constrains the

density change. At Long Valley caldera, CA, for example, the combination of gravity and deformation observations allowed the inference of an intrusion of volatile-rich magma as the source of inflation in the period 1982–1999 [73].

For monitoring purposes, standard campaign-style gravity observations are of limited use, although such repeated measurements are vital for the types of research noted above. There is significant monitoring potential with continuous networks of gravimeters [74]. An example is the continuous gravity network at Mt. Etna, Italy. Continuous gravity observations at the start of the 2002–2003 eruption, showing a reduced gravity decrease followed by recovery over a few hours, have been interpreted as indicating the initial opening of dry fractures that were subsequently filled by magma [75]. Although seismicity commenced several hours before the beginning of the gravity change, the detection of the opening of fractures would certainly be a key part of identifying the likely onset of an eruption.

Volcanic Gas

Detection of volcanic gas is an early indicator of magma ascent that can provide clues to speed of ascent, magma chemistry and explosivity, volume of intruding magma, and the state of the hydrothermal system at a volcano. As magma ascends in the Earth's crust, the decrease in confining pressure leads to exsolution of volatiles from the melt. Additional volatiles can also be released from the existing hydrothermal system as newly emplaced magma heats the surrounding rock. The resulting gases and fluids work their way toward the surface through fractures and can be observed through soil monitoring and space-, air-, and ground-based monitoring of fumaroles and gaseous plumes.

Interpreting the emission rates and compositions of volcanic gases to evaluate magmatic unrest is a challenging task. The compositions of volatiles released from magma vary with tectonic setting, melt composition, and pressure, but the most common gases include, in order of decreasing abundance, water, carbon dioxide (CO_2), sulfur dioxide (SO_2), and halogens. Arc volcanoes often host hydrothermal systems, crater lakes, or can be covered in snow and ice. For these reasons, volcanic gases emitted from arc volcanoes often have chemical reactions with water and other compounds on their way to the surface, which change their chemical form. For example, hydrolysis reactions can change SO_2 into H_2S and native sulfur [76]. This characteristic makes interpreting SO_2 emissions difficult at volcanoes with active hydrothermal systems (e.g., [77]). In contrast, because CO_2 exsolves from magma before other chemical species and because its chemical form is relatively stable as it ascends, its detection at the surface can be an early indicator that magma is ascending and accumulating beneath a volcano [78, 79]. The later appearance of increased SO_2 may then indicate continued magma ascent or drying out of the hydrothermal system. Given these complexities, it is necessary to consider the ratios of CO_2, SO_2, and H_2S to interpret the presence of these gases correctly. An additional complication in the interpretation of gas emissions is that most active volcanoes emit low levels of gas passively; therefore, background monitoring of quiescent time periods must be established before gas emissions associated with unrest can be interpreted correctly. Several strategies must be employed to measure the complete suite of emitted volcanic gases, including both continuous and episodic ground and/or aircraft-based gas monitoring of fumaroles, satellite monitoring, and measurement of CO_2 flux through soils (see [80–82] for reviews of monitoring techniques).

Visual and Thermal Remote Sensing

Recent advances in satellite monitoring and ground-based and airborne remote sensing capabilities have revolutionized volcano monitoring and ash cloud tracking and provided a new perspective for understanding eruption dynamics. In the case of remote volcanoes that lack local seismic and GPS monitoring, satellite data often provide the only data stream documenting unrest and eruption. Data from satellites operated by the various international space agencies are primarily used for weather and climate research and forecasting, but are also used for volcano monitoring tasks including detecting and measuring anomalous thermal emissions, tracking ash clouds, and making visual observations. Rather than describe a complete

list of satellite-based tools used in volcano studies, here the applicability of several satellite systems commonly used at US volcano observatories are highlighted, and readers are referred to more detailed reviews in the literature [82–84]. Some land- and aircraft-based imaging techniques are also briefly described.

When magma intrudes into a volcano, heat flow increases at the Earth's surface, resulting in hot fumaroles and fractures and melting of snow and ice. Thus, thermal remote sensing measurements can provide an early indicator of volcanic unrest in addition to defining the existence and spatial extent of lava flows, domes, and pyroclastic flows. The Moderate Resolution Imaging Spectroradiometer (MODIS) sensors on NASA research satellites and Geostationary Operational Environmental Satellite (GOES) and Advanced Very High Resolution Radiometer (AVHRR) sensors on NOAA satellites provide frequent, low resolution (~1 km pixel) data. Mid-infrared data (3.5–4 μm wavelength) from these sensors are used to study the extent and temporal development of lava and pyroclastic flows and to estimate effusion rates and thermal flux associated with an eruption.

Thermal infrared data from satellites (8–14 μm wavelength) are used to detect and track volcanic ash and gas clouds in the atmosphere, complementing ground- and satellite-based radar measurements of ash clouds. By comparing the brightness temperature difference in two different frequency bands, 11 and 12 μm, clouds containing ash can be discriminated from those containing only water using the Brightness Temperature Difference Method [85]. This method was used for tracking ash clouds from many eruptions (e.g., [86]) and has improved our understanding of the global atmospheric effects of large eruptions. The use of this technique to detect and measure volcanic ash clouds is limited by temporal coverage, atmospheric and cloud water content, tephra particle size, and thermal contrast between the cloud and the surface beneath it. Timing and intervals between images are dependent on satellite position and global location. For northern Pacific volcanoes, for example, GOES data are available every 15 min, while AVHRR data are available 1–12 times per day depending on specific location [87]. Landsat TM and ETM+, Advanced Spaceborne Thermal Emission Reflection Radiometer (ASTER), and other high-resolution sensors provide a complementary dataset to these sensors. Although these data cannot be obtained in real time and images are less frequent, they provide high spatial resolution (15–90 m or better) for detailed visual and thermal observations. A recent compilation of capabilities to detect and measure volcanic clouds can be found in a European Space Agency report [88].

Thermal monitoring is not limited to satellite sensors. As an example of a sensor that can be hand held or mounted on an aircraft or tripod, the Forward Looking Infrared Radiometer (FLIR) camera is highlighted. FLIR surveys and installations often record simultaneous visual and infrared images or movies, providing maps of temperature distributions in ash clouds and on land surfaces. These images can be used to detect fumaroles, map pyroclastic flow deposits, and define spatial extent and structure of lava flows and domes. Frequently, gas emissions visually obscure volcanic activity, but FLIR thermal images can "see" through some gas, as demonstrated at volcanoes such as Mount St. Helens and Augustine [89, 90]. FLIR data have also been used successfully to characterize individual explosion characteristics at Stromboli Volcano [91].

Satellite remote sensing data are used for more than thermal imaging. Ultraviolet spectrometers, such as Total Ozone Mapping Spectrometers (TOMS) and the newer Ozone Monitoring Instrument (OMI) sensor operated by NOAA, can be used to map paths and concentrations of SO_2 clouds emitted from volcanoes by UV absorption of SO_2 in the atmosphere [92, 93]. Thanks to improved capabilities from the OMI sensor, scientists are better able to detect precursory SO_2 emissions in the atmosphere, to quantify eruptive SO_2 more accurately, and to track the SO_2 for longer periods of time. Thus, OMI data provide early indicators of magmatic unrest in addition to improving our understanding of eruption dynamics and hazard to aviation.

The remote sensing capabilities described here will undoubtedly evolve rapidly in the coming decade as technologies continue to advance and new satellites are launched. New remote monitoring tools will continue to emerge as well. For example, recent technical advances permit the measurement of volcanic lightning, both from ground and satellite sensors, to detect and study large ash clouds [94]. Improved resolution in digital cameras has permitted scientists to construct spectacular three-dimensional models of lava dome growth using aerophotogrammetric techniques [95].

Infrasound

Infrasound is the subaudible (<20 Hz) range of sound waves. Infrasound observations are made with either commercial or custom-built low-frequency microphones. These can be deployed individually or, more commonly, as arrays. Array data can be stacked to diminish noise, which can be significant in these data, and can be used to determine the direction to the source. Infrasound disturbances can regularly be detected up to a few hundred kilometers from a volcano during eruptions. The use of infrasound for volcano monitoring was sufficiently rare at the beginning of the twenty-first century for the technique to be absent from some past reviews of volcano monitoring techniques (e.g., [12]), but now such observations are being made at dozens of volcanoes.

Infrasound observations are of great value for monitoring and quantifying eruptions, and are potentially useful for eruption prediction, especially in open-vent systems [96]. Unlike seismic observations, where variations in materials and structure in the Earth's crust along the seismic ray path have a strong influence on the recorded wavefield, the atmosphere alters infrasonic airwaves relatively little at high frequencies [97]. This advantage makes infrasound quite useful for studying eruption dynamics and for quantitative comparisons of eruptions among different volcanoes [98]. Another advantage is that it is very difficult to distinguish between earthquakes at very shallow depths versus earthquakes directly associated with surface explosions using seismic data alone, but in many cases the absence or presence of an infrasound signal can serve to distinguish between the two possibilities and thus provide direct evidence that magmatically driven activity has reached the surface.

Infrasound data provide a new and unique perspective on the dynamics of volcanic eruption columns. Comparing ratios of seismic and acoustic energy between discrete explosions can offer compelling evidence for gas distributions in the magma column and eruption violence. Matoza et al. [99] have used infrasound data to investigate the spectra of volcanic jets and showed that they are similar to noise from aircraft jet engines. Infrasound has also been used to estimate the velocity of material ejected from volcanic vents during eruptions [100]. Observations of this sort could potentially be helpful in estimating ash output at erupting volcanoes when they cannot be observed directly.

Impacts

The literature on the impacts of volcanic eruptions is vast, so only brief and rather general information about selected topics is presented in this entry. The focus is on key primary and secondary volcanic hazards, and in particular those that have relatively immediate and direct impact. As a result, topics such as climatic effects are not covered here.

Lava Flows, Pyroclastics, and Tephra

Primary volcanic hazards that are direct eruption products can be categorized somewhat generally into lava flows, pyroclastic ejecta, and tephra. Lava is molten rock, and surface lava flows can travel many kilometers, or tens of kilometers in the case of basaltic tube-fed flows [101]. Their impact, though, will generally be spatially limited for any single eruption, although cases such as the decades-long rift eruption of Kīlauea Volcano are exceptions. Basaltic lava flows cause direct damage to the natural environment and human infrastructure, as well as igniting fires and/or touching off explosions that result in further damage. Fatalities due to basaltic lava flows are typically minimal, however, due to their modest flow rates, at least on terrain that is not particularly steep. Andesitic-dacitic stratovolcanoes, such as Mount St. Helens, grow blocky and viscous lava domes. Although generally limited in spatial extent to only the volcanic crater and its drainages, these domes can become unstable and fail, producing ash fall and hot block and ash flows down slope.

In contrast to lava flows, pyroclastic flows and surges (dense and dilute solid-gas mixtures, respectively, with a range of possible temperatures) travel at great speed, 10s of meters per second (m/s) to about 150 m/s (over 500 km/h), and large flows can travel 50 km or more from their source vents. The potential for destruction and death is summarily greater. For example, in the 1990s, deaths due to pyroclastic flows comprised the vast majority of directly caused mortality by volcanic eruptions [102].

Tephra is a general term encompassing various types of pyroclastic ejecta that are typically classified

according to size, including blocks and bombs, lapilli, ash, and dust [7]. The larger fragments follow ballistic trajectories, whereas smaller particles can remain suspended in the atmosphere for some time (minutes to weeks) before falling to the surface. Tephra accumulations can amount to tens of centimeters to a few meters at distances of tens to hundreds of kilometers for very large eruptions [103]. Lockwood and Hazlett [7] point out that ballistic fragments are produced in greater abundance by smaller explosive eruptions. Annen and Wagner [102] report a similar number of deaths due to collapse of ash-covered roofs as due to pyroclastic flows and surges in the 1990s.

Lahars

Although a less familiar term to many, these mudflows or debris flows originating from volcanoes can have an enormous impact and cause many fatalities. Lahars can be generated directly and immediately by pyroclastic flows, or in a delayed manner upon collapse of volcanic deposits (for example, due to very heavy rainfall) or due to a lake breakout [104]. Flow rates are generally slower compared to pyroclastic flows, less than 10 m per second except on steep slopes.

A prime example of a lahar is from the eruption of Nevado del Ruiz, Colombia, in 1985. When the volcano erupted violently on the night of November 13, a massive lahar was initiated when pyroclastic flows caused massive melting of snow and glacial ice on the volcano. The lahar swept through and buried the town of Armero nearly 75 km away, resulting in more than 20,000 fatalities [105]. Sadly, the potential for a lahar from Nevado del Ruiz striking Armero had been well documented beforehand – mudflows from Nevado del Ruiz eruptions in 1595 and 1845 buried the same area [106]. Destructive lahars were also produced by the 1980 eruption of Mount St. Helens and the massive 1991 eruption of Mt. Pinatubo, Philippines.

Landslides, Lateral Blasts, and Tsunamis

Weaknesses in the interiors of volcanoes (fracturing and poorly consolidated material, hydrothermal alteration, etc.) leaves them prone to major landslides, which here is meant to encompass also avalanches and sector collapses. Avalanches are common on composite volcanoes, aka stratovolcanoes [7]. Some avalanches are triggered by the shaking from an earthquake. A landslide that grows and becomes more chaotic as it descends is termed a debris avalanche. At a larger scale, a sector collapse is the breaking away of a wedge-shaped flank of a volcano. When combined with an eruption, it can produce a lateral blast. Sector collapses are surprisingly common, at least on a geologic time scale. A number of Hawaii's volcanoes have experienced sector collapses, for example. The 1980 Mount St. Helens eruption began with a sector collapse producing a large debris avalanche, followed by a lateral blast and the initiation of a vertical eruption column. Lateral blasts are somewhat rare, but others have occurred at Arenal Volcano, Costa Rica [107] and Bezymianny Volcano, Russia [108]. For island volcanoes, there is the potential for landslides or especially sector collapses to produce a tsunami. The famous 1883 eruption of Krakatoa was accompanied by repeated tsunamis reportedly as high as 30–40 m around the Sunda Strait, killing tens of thousands of people [11]. Similarly, partial collapse of a volcano neighboring Unzen in 1792 generated a giant tsunami reaching heights of 60 m that caused on the order of 15,000 fatalities [7].

Volcanic Gases

As described above, volcanic gases are important indicators of magma transport, but they also represent a critical hazard. For example, although CO_2 is a significant component of the air we breathe regularly, this odorless, colorless gas is lethal at high concentrations. A tragedy involving CO_2 occurred in 1986 at Lake Nyos, Cameroon, a crater lake formed ~400 years ago. Dissolved CO_2 accumulated in the lake and was released in a discrete event, killing all living things within a 25 km radius, including 1,700 people [109]. In the United States, CO_2 emissions at Mammoth Mountain, CA, led to the deaths of four people between 1998 and 2006. CO_2 is far from the only harmful volcanic gas, however. For example, vog, a form of air pollution resulting from emissions of SO_2 and other volcanic gases, plagues the Island of Hawaii, as the actively erupting Kīlauea Volcano is a prolific producer of several gas species [110].

Eruption Forecasting: Strategies and Challenges

Earthquakes have been recognized to herald volcanic eruptions throughout history. For example, prior to

the well-known eruption of Mount Usu, Japan in 1663, earthquake ground shaking caused local residents to evacuate. Eruption forecasting became a science with the advent of real-time seismic monitoring capabilities. Although earthquake observations still form the backbone of eruption forecasting, they are now complemented by more sophisticated analyses of seismic data and data from many other disciplines. Geodesy, gas chemistry analysis, satellite remote sensing, visual observations and high-resolution photography, petrography, and geochemistry all provide critically important indicators of the state of a magmatic system. Detectable manifestations of magma ascent vary widely between volcanoes based on several factors including the magma chemistry and crystallinity, the physical state of the volcano's conduit system and surrounding crust, the tectonic setting of the volcano, and time since last eruption. Thus, no one-size-fits-all forecasting approach exists. Modern volcano observatories are dynamic organizations that integrate a variety of data streams with knowledge of volcanic history to evaluate the state of unrest and potential future activity at a volcano.

In the United States volcano observatories use a color code alert level system to describe the state of volcano unrest [111]. One of the biggest challenges in forecasting volcanic eruptions and applying this alert system is assessing the time scale over which an eruption might occur. Here, the terms "forecast" and "prediction" as defined by [112] are used. Long-term forecasts, which address hazards on time scales of decades and centuries, are based on geological mapping of volcanoes and their deposits. Correct forecasts that describe likely hazards on time scales of hours to months based on geophysical and gas monitoring are common. However, reliable short-term predictions which specify the time and size of eruptions are difficult and fraught with complexities, not only in correctly interpreting the physical processes responsible for observations but in the delicacies of communicating with the emergency managers who coordinate societal response.

Almost all volcanic eruptions have earthquakes as precursors, but seismological response to magma ascent can vary significantly in character and in magnitude between volcanic systems. Although many eruptions of VEI 3 or smaller at frequently active volcanoes have only small earthquakes (magnitude (M) less than ~2.5) that may not be noticed without local seismic monitoring (e.g., [113]), large eruptions at volcanoes that erupt infrequently can have large earthquakes. For example, the 1980 eruption of Mount St. Helens, the 1991 eruption of Pinatubo Volcano, and the 2008 eruption of Kasatochi Volcano all had associated earthquakes of M5.1 or greater [114]. The cataclysmic rhyolitic eruption of Novarupta in 1912, the largest eruption of the twentieth century, was associated with a staggering nine M6.0 and larger earthquakes [115].

Earthquakes at volcanoes often have unique characteristics that are not observed in purely tectonic systems, like the San Andreas Fault in California. The most obvious difference is that earthquakes at volcanoes tend to occur in swarms of many small earthquakes of similar size, as noted above. Unlike on tectonic faults, these earthquakes tend to increase in magnitude and frequency of occurrence with time before an eruption as magmatically driven pressure increases in the Earth's crust. The time history of seismicity is therefore critical in eruption forecasting. Earthquakes at volcanoes can also have distinctive frequency characteristics. LP and VLP earthquakes and volcanic tremor, for example, reflect fluid movement in the Earth's crust (see the "Seismology" section). Kasatochi Volcano, in the central Aleutian Islands, displayed the classic seismological eruption precursor sequence in 2008 [114]. In the 48 h prior to eruption, the rate and magnitude of earthquakes gradually increased, reflecting pressurization in the Earth's crust. In the 2 h prior to eruption and shortly after a M5.8 earthquake, strong volcanic tremor was observed, indicating that the earthquake likely increased permeability in the crust sufficiently for volatiles and magma to ascend rapidly.

Advanced analyses of seismic data, including some techniques described in the "Seismology" section, are actively being explored for eruption forecasting applications. For example, in some situations, real-time high-precision earthquake locations may be useful. In the 2000 eruption of Miyakejima, earthquakes were observed to migrate laterally as a dike was emplaced [116]. In other situations, however, such as the Long Valley caldera of California, earthquake locations can be misleading, as they reflect geothermally active areas

and the fluid pathways rather than the location of magma itself.

The value of geodesy as an eruption forecasting tool was emphasized to the volcanological community during the 1980 eruption of Mount St. Helens. Prior to eruption, scientists at the Cascades Volcano Observatory documented a growing bulge on the north flank of the mountain. The eruption began as the bulge failed in an earthquake-induced landslide. The surprising and deadly horizontal blast of pyroclastic material that resulted highlighted the importance of deformation data in determining not only volume of magma intruded into volcanic edifices, but also eruption style and potential edifice collapse. Development of continuous GPS technologies in the late twentieth century made near-real-time deformation monitoring a reality. Later development of InSAR methods further improved the role of geodesy in volcano monitoring by providing spatially complete snapshots of volcano deformation. GPS is now a common component of volcano monitoring networks, although it is still significantly less common than seismic monitoring [117]. Dzurisin [118] describes monitoring strategies in which deformation data may be useful in making longer-term forecasts than are typically possible with seismic data alone.

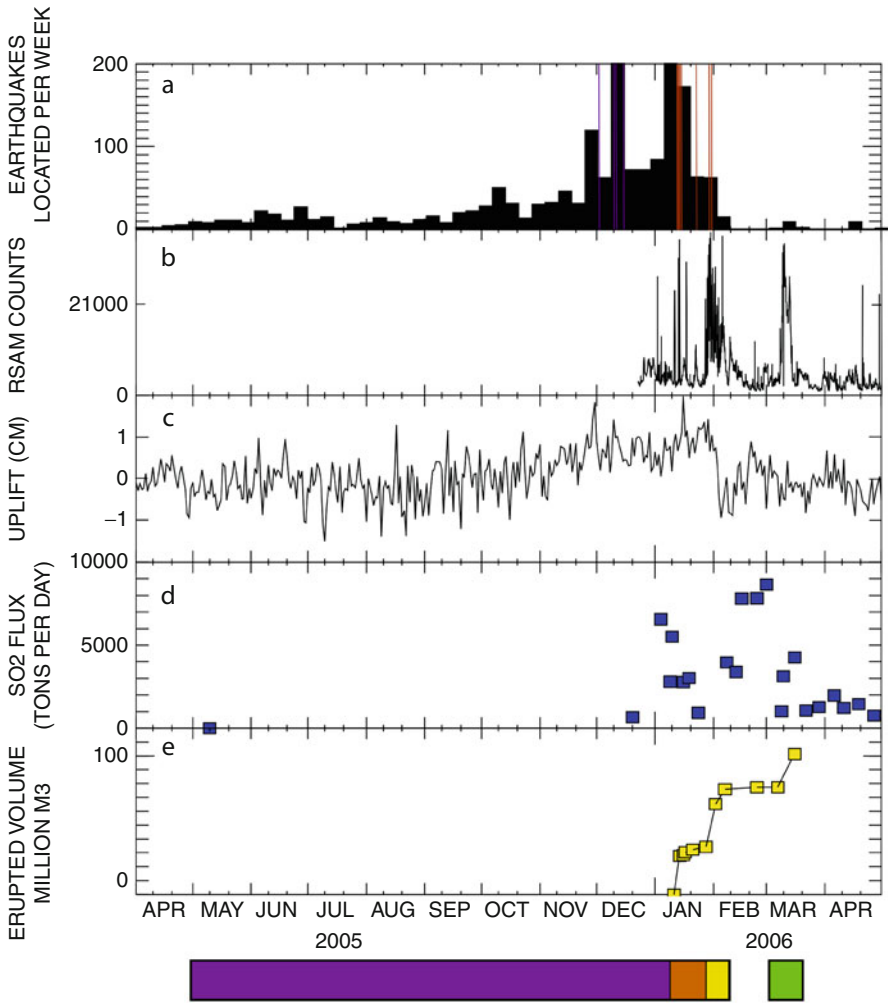
Although seismic and geodetic monitoring are the most common data streams used in eruption forecasting, many other disciplines provide highly valuable information that can be critically important. For example, explosive eruptions of Bezymianny Volcano (Kamchatka), which generally begin with growth of a new lava dome that subsequently becomes unstable and fails, have been forecasted successfully based solely on thermal satellite data by KVERT (the Kamchatkan Volcanic Eruption Response Team). During the preeruptive stage of the 2009 eruption of Redoubt Volcano, Alaska, CO_2 levels provided one of the most conclusive early indicators that the unrest would advance to full eruption, rather than resulting in a stalled intrusion. Geochemical and physical tephra analyses often provide the first indicator that a juvenile magma is involved in an eruption, a critical indicator to understanding the size and potential explosivity of eruptions.

The challenge in interpreting geologic and geophysical indicators of eruption lies in determining the time scale and type of eruption, or even whether the magmatic activity will result in eruption at all. Earthquake swarms, for example, are common in volcanic and geothermal environments, and at least 2/3 and possibly as many as 9/10 of these do not result in eruption [119]. As magmas ascend, volatile exsolution and crystallization reduce overpressure and buoyancy in the magma body. Thus, most ascending magma bodies stall long before reaching the Earth's surface. Rejuvenation of a stalled magma body often involves an external trigger. Magma can be remobilized by many processes including accumulation of volatiles within the body, interaction of magma with water, a decrease in regional confining stress, or infusions of additional gas-rich magma from depth.

Eruption Forecasting: Case Studies

To demonstrate the variability in eruption precursors, two contrasting eruption case studies managed by the Alaska Volcano Observatory (AVO) are described: the 2006 eruption of Augustine Volcano and the 2008 eruption of Okmok Volcano. Both volcanoes were seismically and geodetically monitored, geologically mapped, and had documented historical eruptions. While the Augustine eruption exemplifies a successful eruption response, the Okmok eruption demonstrates how even thorough volcano monitoring does not guarantee that an eruption can be forecast.

The unrest preceding the 2006 eruption of Augustine Volcano, an andesitic stratovolcano located in the lower Cook Inlet of Alaska, proceeded in a "textbook" manner (see Power et al. [120] for a detailed review of the eruption). Scientists had observed this volcano erupt twice previously in 1976 and 1986, providing a template for interpreting unrest. Conveniently for the volcano observatory staff, the volcano closely followed the 1986 template. An increase in earthquake rate was first noted in late April 2005 (Fig. 3; [121]). Subsequently in fall 2005, airborne gas monitoring revealed that SO_2 output increased [122] and GPS monitoring showed that the volcano began to inflate [53]. These data together strongly suggested that magma was ascending beneath and accumulating within the volcanic conduit. Thus, AVO raised the color code for level of concern from Green to Yellow in late November 2005 (see Neal et al. [123] for



Volcanoes, Observations and Impact. Figure 3

Time history of the 2006 eruption of Augustine volcano, Alaska (From [121] and references therein). (**a**) Number of earthquakes located per week. *Purple* and *red lines* indicate phreatic and magmatic explosions, respectively; (**b**) hourly RSAM (reduced seismic amplitude measurement) from station AU13, indicating overall level of seismicity and tremor; (**c**) Uplift relative to GPS stations A59 and AV02; (**d**) SO_2 flux; and (**e**) erupted volume

a detailed chronology of color codes). On December 2, 2005, the first small phreatic (steam) explosion occurred, presumably as ascending magma interacted with water in the conduit. In response to this event, the color code was raised to Orange. Explosions of this type are often interpreted as "vent clearing" events, as they open the pathway for new magma to ascend. In the following month, anomalous activity waxed and waned, but continued to increase gradually in severity, until January 11, 2006, when the volcano had its first significant ash-rich explosion of magma, and the color code was raised to Red. This phase of the eruption continued with intermittent explosions for 2 weeks. Subsequently, the volcano effused small amounts of ash nearly continuously for two additional weeks, through early February 2006. The eruption gradually transitioned to effusive lava dome growth as magma was degassed sufficiently to permit more passive eruption of andesitic magma. A lava dome grew at the volcano's summit through mid-March 2006 and the volcano subsequently returned to a quiet state. This series of increased unrest, phreatic explosions,

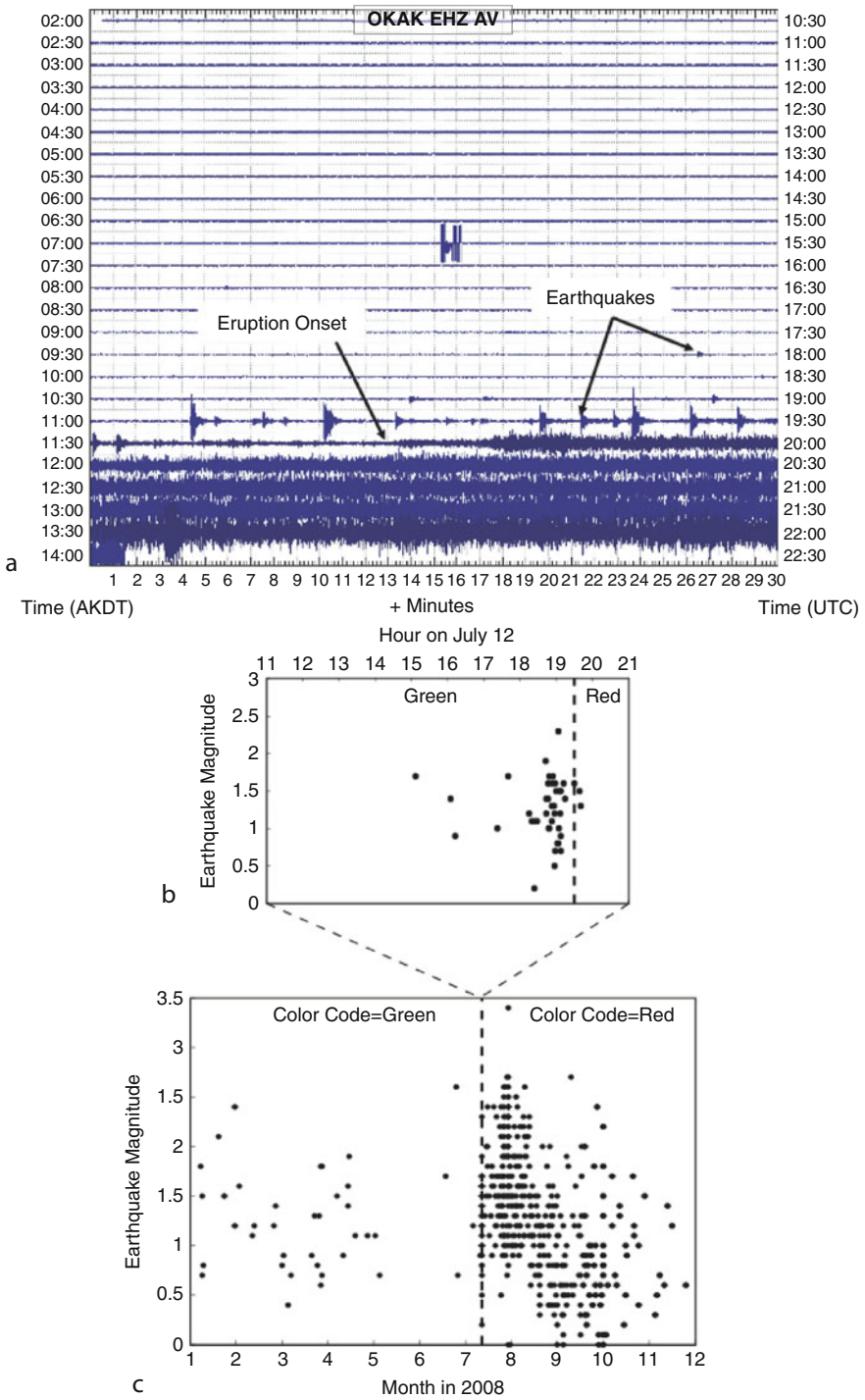
magmatic explosions, and effusion is relatively common in andesitic stratovolcanoes and relatively easy to forecast correctly.

Okmok Volcano, a large caldera located in the eastern Aleutian Islands, had a history of small explosions and basaltic and basaltic-andesite lava flows prior to the 2008 eruption, with the most recent of these in 1997. In the decade between eruptions, earthquakes were rare and volcanic tremor was common enough to be considered background activity for the volcano. GPS and InSAR data revealed that Okmok experienced periods of both inflation and deflation since at least 2000 and was inflating continuously for 6 months prior to eruption [124, 125]. On the morning of July 12, 2008, the volcano was quiet and deformation data revealed no unusual increase in the rate of inflation. However by 11:30 AM AKDT of that day, seismicity ramped up dramatically from quiescence to near constant small earthquakes of $M < 3$ (Fig. 4) over a 60 min interval. Phreato-magmatic eruptive activity followed the first earthquake in the series by less than 2 h. In response, AVO raised the color code from Green directly to Red. Unlike previous effusive lava flow eruptions at Okmok Volcano, the 2008 eruption propelled ash into the atmosphere, affecting north Pacific air traffic for several months. Geologic studies of the eruption deposits reveal that the interaction of magma with a large quantity of groundwater likely drove the eruption to be more explosive and ash-rich than previous eruptions at the volcano [126]. Johnson et al. [127] speculate that prior to the eruption, the shallow open magmatic system was able to degas easily, preventing build up of overpressure and earthquakes. Either a subtle addition of melt or sudden contact with water may have triggered the eruption. This scenario is a worst-case example in terms of forecasting. It is arguably the only eruption at a monitored United States volcano with a well-established seismic network that was not forecast since 1989.

Volcano-Earthquake Interactions

Volcanism and large earthquakes near plate margins are two different manifestations of plate tectonics; thus, over time scales of millennia the processes are closely linked. How earthquakes and volcanoes affect one another over shorter timescales, such as days to months, remains an open and intriguing problem. One frequently asked question in times of crisis is, "can a large earthquake trigger eruption of a nearby volcano?" There are a few compelling examples of eruptions following large earthquakes that suggest a causal relationship between the two, such as the brief eruption of Kīlauea Volcano following the M 7.2 Kalapana Hawaii earthquake in 1975 and the eruption of Puyehue-Cordon Caulle Volcano immediately following the M 9.5 Chile earthquake in 1960. However, it is also true that large earthquakes occur commonly without related eruptions. For example, although the 1964 M9.2 Alaska earthquake occurred in a region with dozens of active volcanoes, the only report of volcanic activity in the subsequent months was a suggestion of increased steaming at Wrangell Volcano, Alaska. Establishing a causal, statistically significant relationship between a large earthquake and a subsequent volcanic eruption is difficult, as it requires an accurate understanding of the probabilities of each event occurring independently. This in turn requires a robust long-term record of eruption and earthquake occurrence, which is rarely available. Studies to date [128–130] suggest that evidence exists for earthquake-triggered eruptions, but it is not a common phenomenon. In the rare occasion where this does occur, there are several possible mechanisms that may explain eruption triggering [131]. Static stress changes in the Earth's crust related to the earthquake may change the confining pressure on a magma reservoir, destabilizing it. Alternatively, high frequency dynamic seismic waves may destabilize the magma by causing bubble nucleation or increasing convection within the reservoir. Finally, violent shaking may cause fractures, landslides, or other changes to the crustal volume surrounding a shallow magma reservoir, leading to eruption.

Large earthquakes commonly affect volcanoes in subtle ways, however. Following the 1992 M 7.3 Landers, California earthquake, volcanic and geothermal areas across the western United States showed an increase in earthquake activity [132]. Since that time, small earthquakes triggered at great distances by surprisingly small amplitude oscillatory seismic waves from large earthquakes ($\leq$0.01 MPa dynamic stress change) have been documented at many areas around the world (see Prejean and Hill [133] for review). Although dynamic triggering of small earthquakes



Volcanoes, Observations and Impact. Figure 4

Seismicity related to the 2008 eruption at Okmok caldera, Alaska, from [125]. (**a**) Twelve hours helicorder plot of seismic data from station OKAK, showing quiescence in the hours prior to eruption. (**b**) Located earthquakes over 10 h near the time of eruption onset. (**c**) Located earthquakes at Okmok by the Alaska Volcano Observatory in 2008. *Colors* indicate color code alert level assigned by the Alaska Volcano Observatory

happens in many environments, volcanoes appear to be particularly susceptible to dynamic earthquake triggering [133]. A range of physical models have been proposed to explain how small amplitude dynamic waves trigger earthquakes, including changing fluid pathways in delicate hydrothermal systems, changing the crustal stress field by disrupting magma chambers, and directly exceeding the frictional strength of faults [134]. Most models require very high pore-fluid pressures in the volume where the triggered earthquakes occur.

Volcanoes and Sustainability

The focus of this section is on three direct and immediate connections between volcanoes and sustainability. One is the societal hazard factor – as with other geologic hazards, population growth puts more and more people at risk for possible volcanic impacts. A second, related impact is eruption effects on aviation, which received global attention in 2010 due to the eruption of Eyjafjallajökull Volcano in Iceland. The third is magmatic activity as a source of geothermal energy. Areas being tapped for commercial-scale geothermal energy are in regions of magmatic activity if not actually on the flanks of a volcano (e.g., the Puna Geothermal Venture on Kilauea). Longer-term connections between volcanoes and sustainability, including climatic effects and connections to ore deposits, are not covered here but are discussed in numerous sources (e.g., [135, 136]).

Volcanoes and Human Population

About 10% of the world's population lives on or near active volcanoes [137], and this percentage is steadily increasing with time. Lockwood and Hazlett [7] estimate that more than 100 million people live in areas near calderas that have been subjected to pyroclastic flows. There have been a number of large eruptions in the past several centuries, but nothing rivaling the very large explosive volcanic eruptions ("super volcanoes") that are present in the geologic record [138], so in historic times, humans have not experienced the full impact that volcanic eruptions can produce. However, the number of eruptions causing fatalities has steadily increased each century since the 1500s, which Simkin et al. [139] attribute to increased global population, as opposed to an increase in the frequency of eruptions.

Quantifying the hazard from volcanic eruptions is certainly a challenge, especially considering the shortness of the historic record in comparison to the frequency of occurrence of large to very large eruptions (centuries to many millennia). Interestingly, eruption size on the VEI scale [8] follows a roughly linear size-log frequency distribution, similar to the Gutenberg-Richter relation for earthquakes. Simkin and Siebert [11] found that VEI 6 eruptions occur about once or twice a century, VEI 5 about once per decade, VEI 4 once or twice a year, etc. Ewert and Harpel [140] combined the global distributions of population and volcanoes to derive a Volcano Population Index (VPI), and applied it to Central America. VPI quantifies the population within 5 (VPI5) or 10 (VPI10) km of a volcanic system, corresponding to eruptions of VEI 2–4. They found that roughly 2.5 million people were at risk as measured by VPI10, and also suggested that only Indonesia might have a larger exposed population than Central America. As in the case of earthquake and tsunami hazards, population centers in harm's way of a possible future eruption are certainly not going to relocate. Mitigation, warning strategies and systems, and education efforts are therefore essential.

Volcanoes and Aircraft

The prolonged eruption of Eyjafjallajökull in Iceland in 2010 is a prime example of the impact volcanic eruptions can have on air travel. That disruption cost airlines billions of dollars [141], and certainly caused additional losses to businesses and people impacted by the shut down in air travel. The most critical hazard is the intake of ash into jet engines, which can severely damage engine parts and potentially cause engine failure. For example, KLM Flight 867, with 245 passengers and crew on board, encountered the ash cloud from the eruption of Alaska's Redoubt Volcano on December 15, 1989, causing all four engines to shutdown. The airplane plunged nearly 15,000 ft before the pilots were able to restart the engines and ultimately land safely, albeit with tens of millions of dollars of damage to the airplane. In the 1980s through the mid 1990s, about one airplane per year experienced damage from volcanic ash on North Pacific air routes [142]. The USGS, along with

Alaskan state agencies, established the Alaska Volcano Observatory in 1988 in large part to address the volcanic ash-aircraft problem in that region.

Assessing and monitoring the ash hazard to aircraft requires the application of essentially all the tools available. This includes forecasting, monitoring, and prediction efforts prior to an eruption and 24/7 geophysical monitoring and remote sensing to track eruptive activity and the spread of ash clouds during an eruption. Because volcanic ash from large eruptions can remain in the atmosphere at flight altitudes for weeks or longer, circling the globe multiple times, addressing the hazard requires international cooperation of air traffic controllers, scientists, and weather agencies. As air passenger and cargo air traffic continues to increase globally, the risk will escalate.

Volcanoes and Geothermal Energy

Volcanoes can provide part of the solution to global energy needs, as heat generated inside the Earth can be harnessed for geothermal energy. Although volcanic activity is not required for geothermal energy production in enhanced geothermal systems (EGS), such as that at Soultz-sous-Forets, France, volcanic activity brings heat nearer the Earth's surface, potentially lowering the cost of energy production. The United States is currently the largest producer of geothermal energy [143], although many other countries are developing their own resources. Successful geothermal energy plants located on active volcanoes or above magma chambers are currently in operation in the United States at Kīlauea Volcano in Hawaii, the Long Valley caldera, California, and many other sites in California, Nevada, and Idaho, although in total geothermal sources provided less than a half a percent of the electricity in the US in 2009 [144]. The potential to produce geothermal energy at additional volcanoes in Alaska and the Pacific Northwest is being explored aggressively. Most current geothermal energy operations involve tapping hot ground water. In contrast, the Geysers, California – the largest geothermal field in the United States – is a dry steam field that taps superheated steam.

A potential risk associated with geothermal power is triggering earthquakes. Generally, earthquakes associated with geothermal production are small ($M < 2$), but occasional larger earthquakes have disrupted communities, most notably in Basel, Switzerland. A second less common but significant risk to geothermal energy production at active volcanoes in Hawaii and Alaska involves eruptions disrupting the energy supplies and potentially destroying infrastructure.

Future Directions

Technological developments to improve and expand observational capabilities are likely to have the greatest impact on the development of volcano studies in the future [145]. An example is the use of autonomous sensor networks [146, 147]. The combination of ease of deployment and networked communication makes such systems extremely appealing for volcano monitoring. Deployment of a constellation of radar satellites with wavelengths that are tuned to "see" through vegetative cover could cut InSAR repeat observation cycles to intervals of days, allowing for high spatial resolution observations of deformation on a time scale of use for eruption forecasting. One can imagine a day when radar satellites detect significant deformation of a volcano anywhere in the world (on land), an autonomous sensor network is deployed rapidly (e.g., [148]), and a team of volcanologists forecasts the activity as the episode of unrest unfolds, comparable to severe storm forecasting.

In closing, the concern that population growth and expanding commercialism (especially air travel) will inexorably lead to increasing impact of volcanic activity on humans, either directly or indirectly, is reiterated. Although volcanoes certainly cannot be controlled, monitoring technologies and strategies for forecasting and predicting their behavior can be effective. It therefore seems clear that increased efforts in volcano monitoring are required to mitigate future risks.

Bibliography

1. Oppenheimer C (2003) Climatic, environmental and human consequences of the largest known historic eruption: Tambora volcano (Indonesia) 1815. Prog Phys Geogr 27:230–259
2. Stothers RB (1984) The great Tambora eruption in 1815 and its aftermath. Science 224:1191–1198
3. Rose WI, Chesner CA (1987) Dispersal of ash in the great Toba eruption, 75 ka. Geology 15:13–917
4. Hough S (2009) Predicting the unpredictable: the tumultuous science of earthquake prediction. Princeton University Press, Princeton

5. White RA, McCausland WA, Lockhart AB (2011) Volcano monitoring: keep it simple – less can be more during volcano crises; 25 years of VDAP experience. Seism Res Lett 82:330
6. Le Bas MJ, Le Maitre RW, Streckeisen A, Zanettin B (1986) A chemical classification of volcanic rocks based on the total alkali-silica diagram. J Petrol 27:745–750
7. Lockwood JP, Hazlett W (2010) Volcanoes – global perspectives. Wiley-Blackwell, Hoboken
8. Newhall CG, Self S (1982) The volcanic explosivity index (VEI): an estimate of explosive magnitude for historical volcanism. J Geophys Res 87:1231–1238
9. Pyle DM (2000) Sizes of volcanic eruptions. In: Sigurdsson H, Houghton BF, McNutt SR, Rymer H (eds) Encyclopedia of volcanoes. Academic Press, San Diego
10. Decker RW (1986) Forecasting volcanic eruptions. Ann Rev Earth Planet Sci 14:267–291
11. Simkin T, Siebert L (1994) Volcanoes of the world. Geoscience, Tucson
12. McNutt SR (1996) Seismic monitoring of volcanoes: a review of the state-of-the-art and recent trends. In: Scarpa R, Tilling R (eds) Monitoring and mitigation of volcano hazards. Springer, Berlin
13. McNutt SR (2000) Seismic monitoring. In: Sigurdsson H, Houghton BF, McNutt SR, Rymer H (eds) Encyclopedia of volcanoes. Academic Press, San Diego
14. Lahr JC, Chouet BA, Stephens CD, Power JA, Page RA (1994) Earthquake classification, location, and error analysis in a volcanic environment: implications for the magmatic system of the 1989–1990 eruptions at Redoubt volcano, Alaska. J Volcanol Geotherm Res 62:137–151
15. Hill DP, Dawson P, Johnston MJS, Pitt AM, Biasi G, Smith K (2002) Very-long-period volcanic earthquakes beneath Mammoth Mountain, California. Geophys Res Lett 29:1370. doi:10.1029/2002GL014833
16. Hotovec AJ, Prejean SG, Vidale JE, Gomberg J (in press) Strongly gliding harmonic tremor during the 2009 eruption of Redoubt volcano. J Volcanol Geotherm Res
17. Chouet B (1985) Excitation of a buried magmatic pipe: a seismic source model for volcanic tremor. J Geophys Res 90:1881–1893
18. Julian B (1994) Volcanic tremor: nonlinear excitation by fluid flow. J Geophys Res 99:11859–11877
19. White RA (1996) Precursory deep long-period earthquakes at Mount Pinatubo: spatial-temporal link to a basaltic trigger. In: Newhall CG, Punongbayan RS (eds) Fire and mud: eruptions and lahars of Mount Pinatubo, Philippines. University of Washington Press, Seattle
20. Power JA, Stihler SD, White RA, Moran SC (2004) Observations of deep long-period (DLP) seismic events beneath Aleutian arc volcanoes; 1989–2002. J Volcanol Geotherm Res 138: 243–26
21. Mavonga T, Zana N, Durrheim RJ (2010) Studies of crustal structure, seismic precursors to volcanic eruptions and earthquake hazard in the eastern provinces of the Democratic Republic of Congo. J Afr Earth Sci 58:623–633. doi:10.1016/j.jafrearsci.2010.08.008, ISSN 1464-343X
22. Harrington RM, Brodsky EE (2007) Volcanic hybrid earthquakes that are brittle-failure events. Geophys Res Lett 34:L06308. doi:10.1029/2006GL028714
23. Kawakatsu H, Ohminato T, Ito H, Kuwahara Y (1992) Broadband seismic observation at the Sakurajima volcano, Japan. Geophys Res Lett 19:1959–1962
24. Kawakatsu H, Ohminato T, Ito H (1994) 10s-period volcanic tremors observed over a wide area in southwestern Japan. Geophys Res Lett 21:1963–1966. doi:10.1029/94GL01683
25. Neuberg J, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli volcano. Geophys Res Lett 21:749–752. doi:10.1029/94GL00377
26. Kaneshima S, Kawakatsu H, Matsubayashi H, Sudo Y, Tsutsui T, Ohminato T, Ito H, Uhira K, Yamasato H, Oikawa J, Takeo M, Iidaka T (1996) Mechanism of phreatic eruptions at Aso volcano inferred from near-field broadband seismic observations. Science 273:642–645
27. Ohminato T, Chouet BA, Dawson P, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath Kilauea volcano. J Geophys Res 103:23839–23862. doi:10.1029/98JB01122
28. Arciniega-Ceballos A, Chouet BA, Dawson P (1999) Very long period signals associated with vulcanian explosions at Popocate´petl volcano, Mexico. Geophys Res Lett 26:3013–3016. doi:10.1029/1999GL005390
29. Legrand D, Kaneshima S, Kawakatsu H (2000) Moment tensor analysis of near-field broadband waveforms observed at Aso volcano, Japan. J Volcanol Geotherm Res 101:155–169. doi:10.1016/S0377-0273(00)00167-0
30. Nishimura T, Kobayashi T, Ohtake M, Sato H, Nakamichi H, Tanaka S, Sato M, Ueki S, Hamaguchi H (2000) Source process of very long period seismic events associated with the 1998 activity of Iwate volcano, northeastern Japan. J Geophys Res 105:19135–19147. doi:10.1029/2000JB900155
31. Rowe CA, Aster RC, Kyle PR, Dibble RR, Schlue JW (2000) Seismic and acoustic observations at Mount Erebus volcano, Ross Island, Antarctica, 1994–1998. J Volcanol Geotherm Res 101:105–128. doi:10.1016/S0377-0273(00)00170-0
32. Kumagai H, Ohminato T, Nakano M, Ooi M, Kubo A, Inoue H, Oikawa J (2001) Very-long-period seismic signals and caldera formation at Miyake Island, Japan. Science 293:687–690. doi:10.1126/science.1062136
33. Almendros J, Chouet B, Dawson PB, Bond T (2002) Identifying elements of the plumbing system beneath Kilauea volcano, Hawaii, from the source locations of very-long-period signals. Geophys J Int 148:303–312
34. Hidayat D, Voight B, Chouet B, Dawson P, Ratdomopurbo A (2002) Source mechanism of very-long-period signals accompanying dome growth activity at Merapi volcano, Indonesia. Geophys Res Lett 29. doi:10.1029/2002GL015013
35. Aster R, Mah S, Kyle P, McIntosh W, Dunbar N, Johnson J, Ruiz M, McNamara S (2003) Very long period oscillations of Mount Erebus volcano. J Geophys Res 108:2522. doi:10.1029/2002JB002101

36. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, Luca GD, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli volcano, Italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108:2019. doi:10.1029/2002JB001919
37. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocate´petl volcano, Mexico, determined from waveform inversions of very long period signals. J Geophys Res 110:B07301. doi:10.1029/2004JB003524
38. Waite GP, Chouet BA, Dawson PB (2008) Eruption dynamics at Mount St. Helens imaged from broadband seismic waveforms: interaction of the shallow magmatic and hydrothermal systems. J Geophys Res 113:B02305. doi:10.1029/2007JB005259
39. Hill DP (1977) A model for earthquake swarms. J Geophys Res 82:1347–1352. doi:10.1029/JB082i008p01347
40. Foulger GR, Julian BR, Pitt AM, Hill DP, Malin P, Shalev E (2003) Three-dimensional crustal structure of Long Valley Caldera, California, and evidence for the migration of CO_2 under Mammoth Mountain. J Geophys Res 108:B3. doi:10.1029/2000JB000041
41. Patanè D, Barberi G, Cocina O, De Gori P, Chiarabba C (2006) Time resolved seismic tomography detects magma intrusions at Mount Etna. Science 313:821–823
42. Titzschkau T, Savage M, Hurst T (2010) Changes in attenuation related to eruptions of Mt. Ruapehu volcano, New Zealand. J Volcanol Geotherm Res 190:168–178
43. De Gori P, Chiarabba C, Giampiccolo E, Martinez-Arevalo C, Patane D (2011) Body wave attenuation heralds incoming eruptions at Mount Etna. Geology 39:503–506
44. Miller V, Savage M (2001) Changes in seismic anisotropy after volcanic eruptions: evidence from Mount Ruapehu. Science 293:2231–2233
45. Patanè D, De Gori P, Chiarabba C, Bonaccorso A (2003) Magma ascent and the pressurization of Mount Etna's volcanic system. Science 299:2061–2063
46. Volti T, Crampin S (2003) A four-year study of shear-wave splitting in Iceland: 2. Temporal changes before earthquakes and volcanic eruptions. In: Nieuwland DA (ed) New insights into structural interpretation and modeling, Geological Society of London, Special Publication 212. Geological Society, London, pp 135–149
47. Musumeci C, Cocina O, De Gori P, Patanè D (2004) Seismological evidence of stress induced by dike injection during the 2001 Mt Etna eruption. Geophys Res Lett 31:L07617. doi:10.1029/2003GL019367
48. Bianco F, Scarfı L, Del Pezzo E, Patanè D (2006) Shear wave splitting changes associated with the 2001 volcanic eruption on Mt. Etna. Geophys J Int 167:959–967
49. Roman DC, Savage MK, Arnold R, Latchman JL, De Angelis S (2011) Analysis and forward modeling of seismic anisotropy during the ongoing eruption of the Soufrière Hills volcano, Montserrat, 1996–2007. J Geophys Res 116:B03201. doi:10.1029/2010JB007667
50. Brenguier F, Shapiro N, Campillo M, Ferrazzini V, Duputel Z, Coutant O, Nercessian A (2008) Towards forecasting volcanic eruptions using seismic noise. Nat Geosci 1:126–130
51. Poland M, Hamburger M, Newman A (2006) The changing shapes of active volcanoes: history, evolution, and future challenges for Volcano Geodesy. J Volcanol Geotherm Res 150:1–13
52. Dzurisin D (2007) Volcano deformation: geodetic monitoring techniques. Springer, Berlin
53. Cervelli PF, Fournier TJ, Freymueller JT, Power JA, Lisowski M, Pauk BA (2010) Geodetic constraints on magma movement and withdrawal during the 2006 eruption of Augustine volcano. In: Power JA, Coombs ML, Freymueller JT (eds) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston, pp 427–452
54. Dow JM, Neilan RE, Rizos C (2009) The International GNSS service in a changing landscape of Global Navigation Satellite Systems. J Geodesy 83:191–198. doi:10.1007/s00190-008-0300-3
55. Massonnet D, Rossi M, Carmona C, Adragna F, Peltzer G, Feigl K, Rabaute T (1993) The displacement field of the Landers earthquake mapped by radar interferometry. Nature 364:138–142
56. Massonnet D, Briole P, Arnaud A (1995) Deflation of Mount Etna monitored by spaceborne radar interferometry. Nature 375:567–570
57. Thatcher W, Massonnet D (1997) Crustal deformation at Long Valley Caldera, eastern California, 1992–1996 inferred from satellite radarinterferometry. Geophys Res Lett 24:2519–2522
58. Wicks C Jr, Thatcher W, Dzurisin D (1998) Migration of fluids Beneath Yellowstone Caldera inferred from satellite radar interferometry. Science 282:458–462
59. Sigmundsson F, Durand P, Massonnet D (1999) Opening of an eruptive fissure and seaward displacement at Piton de la Fournaise volcano measured by RADARSAT satellite radar interferometry. Geophys Res Lett 26:533–536
60. Lu Z, Fatland R, Wyss M, Li S, Eichelberer J, Dean K, Freymueller J (1997) Deformation of New Trident volcano measured by ERS-1 SAR interferometry, Katmai National Park, Alaska. Geophys Res Lett 24:695–698
61. Lu Z, Mann D, Freymueller JT, Meyer DJ (2000) Synthetic aperture radar interferometry of Okmok volcano, Alaska: radar observations. J Geophys Res Solid Earth 105:10791–10806
62. Lu Z, Wicks C, Dzurisin D, Thatcher W, Freymueller JT, McNutt SR, Mann D (2000) Aseismic inflation of Westdahl volcano Alaska, revealed by satellite radar interferometry. Geophys Res Lett 27:1567–1570
63. Lu Z, Wicks C, Power JA, Dzurisin D (2000) Ground deformation associated with the March 1996 earthquake swarm at Akutan volcano Alaska, revealed by satellite radar interferometry. J Geophys Res 105:21483–21495
64. Lu Z, Power JA, McConnell VS, Wicks C, Dzurisin D (2002) Preeruptive inflation and surface interferometric coherence characteristics revealed by satellite radar interferometry at Makushin volcano, Alaska: 1993–2000. J Geophys Res 107:B11

65. Lu Z, Masterlark T, Power J, Dzurisin D, Wicks C (2002) Subsidence at Kiska volcano, Western Aleutians, detected by satellite radar interferometry. Geophys Res Lett 29:18
66. Jonsson S, Zebker K, Cervelli P, Segall P, Garbeil H, Mouginis-Mark P, Rowland S (1999) A shallow-dipping dike fed the 1995 flank eruption at Fernandina volcano, Galapagos, observed by satellite radar interferometry. Geophys Res Lett 26:1077–1080
67. Amelung F, Oppenheimer C, Segall P, Zebker H (2000) Ground deformation near Gada 'Ale volcano, Afar, observed by radar interferometry. Geophys Res Lett 27:3093–3096
68. Pritchard ME, Simons M (2002) A satellite geodetic survey of large-scale deformation of volcanic centres in the central Andes. Nature 418:167–171
69. Goldstein RM, Zebker HA, Werner CL (1988) Satellite radar interferometry – two-dimensional phase unwrapping. Radio Sci 23:713–720
70. Gens R (2003) Two-dimensional phase unwrapping for radar interferometry: developments and new challenges. Int J Remote Sens 24:703–710
71. Sturkell E, Einarsson P, Sigmundsson F, Geirsson H, Olafsson H, Pedersen R, de Zeeuw-van Dalfsen E, Linde AT, Sacks SI, Stefansson R (2006) Volcano geodesy and magma dynamics in Iceland. J Volcanol Geotherm Res 150:14–34
72. Rymer H (1996) Microgravity monitoring. In: Scarpa R, Tilling R (eds) Monitoring and mitigation of volcano hazards. Springer, Berlin
73. Battaglia M, Hill D (2009) Analytical modeling of gravity changes and crustal deformation at volcanoes: the Long Valley Caldera (CA) case study. Tectonophysics 471:45–57
74. Williams-Jones G, Rymer H, Mauri G, Gottsmann J, Poland M, Carbone D (2008) Toward continuous 4D microgravity monitoring of volcanoes. Geophysics 73:WA19–WA28
75. Carbone D, Budettaa G, Greco F, Rymer H (2003) Combined discrete and continuous gravity observations at Mount Etna. J Volcanol Geotherm Res 123:123–135
76. Symonds RB, Gerlach TM, Reed MH (2001) Magmatic gas scrubbing: implications for volcano monitoring. J Volcanol Geotherm Res 108:303–341
77. Doukas MP, Gerlach TM (1995) Sulfur dioxide scrubbing during the 1992 eruption of Crater Peak, Mount Spurr, Alaska. In: Keith T (ed) The 1992 eruptions of Crater Peak Vent, Mount Spurr Volcano, Alaska, U.S. Geological Survey Bulletin B-2139. U.S. G.P.O.: U.S. Dept. of the Interior, US Geological Survey, Washington, DC, pp 47–57
78. Aiuppa A, Moretti R, Federico C, Giudice G, Gurrieri S, Liuzzo M, Papale P, Shinohara H, Valenza M (2007) Forecasting Etna eruptions by real-time observation of volcanic gas composition. Geology 35:1115–1118
79. Werner C, Kelly PJ, Doukas M, Lopez T, Pfeffer M, McGimsey RG, Neal CA (in press) Degassing associated with the 2009 eruption of Redoubt volcano, Alaska. J Volcanol Geotherm Res (Special Issue on the 2009 Redoubt Eruption)
80. Francis P, Horrocks L, Oppenheimer C (2000) Monitoring gases from andesite volcanoes. Philos Trans Math Phys Eng Sci 358:1567–1584
81. Edmonds M (2008) New geochemical insights into volcanic degassing. Philos Trans Math Phys Eng Sci 366:4559–4579
82. Moran SC, Freymueller JT, LaHusen RG, McGee KA, Poland MP, Power JA, Schmidt DA, Schneider DJ, Stephens G, Werner CA, White RA (2008) Instrumentation recommendations for volcano monitoring at US volcanoes under the National Volcano Early Warning System. USGS Scientific Investigations Report 2008–5114
83. Dean KG, Dehn J, Engle K, Izbekov P, Papp K (2002) Operational satellite monitoring of volcanoes at the Alaska Volcano Observatory. In: Harris AJH, Wooster M, Rothery DA (eds) Monitoring volcanic hotspots using thermal remote sensing. Adv Environ Monit Model 1:70–97
84. Mouginis-Mark PJ, Crisp JA, Fink JH (eds) (2000) Remote sensing of active volcanism, AGU Geophysical Monograph 116. American Geophysical Union, Washington, DC
85. Prata J (1989) Observations of volcanic ash clouds in the 10–12 μm window using AVHRR/2 data. Int J Remote Sens 10:751–761
86. Corradini S, Merucci L, Prata AJ, Piscini A (2010) Volcanic ash and SO_2 in the 2008 Kasatochi eruption: retrievals comparison from different IR satellite sensors. J Geophs Res 115:D00L21. doi:10.1029/2009JD013634
87. Schneider DJ, Dean KG, Dehn J, Miller TP, Kirianov VY (2000) Monitoring and analysis of volcanic activity using remote sensing data at the Alaska Volcano Observatory: case study for Kamchatka, Russia, December 1997. In: Mouginis-Mark PJ, Crisp JA, Fink JH (eds) Remote sensing of active volcanism, AGU Geophysical Monograph 116. American Geophysical Union, Washington, DC
88. Zehner E (2010) Monitoring volcanic ash from space. European Space Agency, Noordwijk, p 110
89. Schneider DJ, Vallance JW, Wessels RL, Logan M, Ramsey MS (2008) Use of thermal infrared imaging for monitoring renewed dome growth at Mount St. Helens, 2004. In: Sherrod DR, Scott WE, Stauffer PH (eds) A volcano rekindled; the renewed eruption of Mount St. Helens, 2004–2006, U.S. Geological Survey Professional Paper 1750. U.S. Dept. of the Interior, U.S. Geological Survey, Reston, p 856 and DVD-ROM [http://pubs.usgs.gov/pp/1750/]
90. Wessels RL, Coombs ML, Schneider DJ, Dehn J, Ramsey MS (2010) High-resolution satellite and airborne thermal infrared imaging of the 2006 eruption of Augustine volcano. In: Power JA, Coombs ML, Freymueller JT (eds) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston, pp 527–552
91. Patrick MR, Harris AJL, Ripepe M, Dehn J, Rothery DA, Calvari S (2007) Strombolian explosive styles and source conditions: insights from thermal (FLIR) video. Bull Volcanol 69:769–784
92. Krueger AJ, Schaefer SJ, Krotkov N, Bluth G, Barker S (2000) Ultraviolet remote sensing of volcanic emissions. In: Mouginis-Mark PJ, Crisp JA, Fink JH (eds) Remote sensing of active volcanism, AGU Geophysical Monograph 116. American Geophysical Union, Washington, DC

93. Carn SA, Krueger AJ, Krotkov NA, Yang K, Evans K (2009) Tracking volcanic sulfur dioxide clouds for aviation hazard mitigation. Nat Hazard 51:325–343
94. McNutt SR, Williams ER (2010) Volcanic lightening: global observations and constraints on source mechanisms. Bull Volcanol 72:1153–1167
95. Schilling SP, Thompson RA, Messerich JA, Iwatsubo EY (2008) Use of digital aerophotogrammerty to determine rates of lava dome growth, Mount St. Helens, Washington, 2004–2005. In: Sherrod DR, Scott WE, Stauffer PH (eds) A volcano rekindled; the renewed eruption of Mount St. Helens, 2004–2006, U.S. Geological Survey Professional Paper 1750. U.S. Dept. of the Interior, U.S. Geological Survey, Reston, p 856 and DVD-ROM [http://pubs.usgs.gov/pp/1750/]
96. Garces MA, Iguchi M, Ishihara K, Morrissey M, Sudo Y, Tsutsui T (1999) Infrasonic precursors to a Vulcanian eruption at Sakurajima volcano, Japan. Geophys Res Lett 26:2537–2540
97. Johnson JB (2003) Generation and propagation of infrasonic airwaves from volcanic explosions. J Volcanol Geotherm Res 121:1–14
98. Johnson JB, Aster RC, Ruiz MC, Malone SD, McChesney PJ, Lees JM, Kyle PR (2003) Interpretation and utility of infrasonic records from erupting volcanoes. J Volcanol Geotherm Res 121:15–63
99. Matoza RS, Fee D, Garces MA, Seiner JM, Ramon PA, Hedlin MAH (2009) Infrasonic jet noise from volcanic eruptions. Geophys Res Lett 36. doi:1029/2008GL036486
100. Caplan-Auerbach J, Bellesiles A, Fernandes JK (2010) Estimates of eruption velocity and plume height from infrasonic recordings of the 2006 eruption of Augustine volcano, Alaska. J Volcanol Geotherm Res 189:12–18
101. Blong R (1996) Volcanic hazards risk assessment. In: Scarpa R, Tilling R (eds) Monitoring and mitigation of volcano hazards. Springer, Berlin
102. Annen C, Wagner J-J (2003) The impact of volcanic eruptions during the 1990s. Nat Hazard Rev 4:169–175
103. Hoblitt RP, Miller CD, Scott WE (1987) Volcanic hazards with regard to siting nuclear-power plants in the Pacific Northwest. U.S. Geological Survey Open-File Report 87-297
104. Siebert L (1996) Hazards of large debris avalanches. In: Scarpa R, Tilling R (eds) Monitoring and mitigation of volcano hazards. Springer, Berlin
105. Ewert JW, Murray T, Lockhart A, Miller C (1993) Preventing volcanic catastrophe: the U. S. International Volcano Disaster Assistance Program. Earthq Volcanoes 24:270–291
106. Wright TL, Pierson TC (1992) Living with volcanoes: The U. S. Geological Survey's Volcano Hazards Program, USGS Circular 1973. United States Government Printing Office, Washington, DC
107. Alvarado GE, Soto GJ, Schmincke H-U, Blge LL, Sumita M (2006) The 1968 andesitic lateral blast eruption at Arenal volcano, Costa Rica. J Volcanol Geotherm Res 157:9–33
108. Fisher RV, Heiken G, Hulen J (1998) Volcanoes: crucibles of change. Princeton University Press, Princeton
109. Holloway M (2000) The killing lakes. Sci Am 283:92–99
110. Sutton AJ, Elias T (1993) Volcanic gases create air pollution on the Island of Hawai'i: U.S. Geological Survey. Earthq Volcanoes 24:178–196
111. Gardner CA, Guffanti MC (2006) U.S. Geological Survey's alert notification system for volcanic activity. U.S. Geological Survey Fact Sheet 2006-3139p
112. Swanson DA, Casadevall TJ, Dzurisin D, Holcomb RT, Newhall CG, Malone SD, Weaver CS (1985) Forecasts and predictions of eruptive activity at Mount St. Helens, USA: 1974–1984. Science 3:397–423
113. Power JA, Jolly A, Nye C, Harbin M (2002) A conceptual model of the Mount Spurr magmatic system from seismic and geochemical observations of the 1992 Crater Peak eruption sequence. Bull Volcanol 64:206–218
114. Ruppert NA, Prejean S, Hansen RA (2011) Seismic swarm associated with the 2008 eruption of Kasatochi volcano, Alaska: earthquake locations and source parameters. J Geophys Res 116:B00B07. doi:10.1029/2010JB007435
115. Abe K (1992) Seismicity of the caldera-making eruption of Mount Katmai, Alaska in 1912. Bull Seismol Soc Am 82:175–191
116. Japan Meteorological Agency (JMA) (2000) Recent seismic activity in the Miyakejima and Niijima-Kozushima region, Japan – the largest earthquake swarm ever recorded. Earth Planets Space 52:i–iv
117. Guffanti M, Diefenbach AK, Ewert JW, Ramsey DW, Cervelli PF, Schilling SP (2008) Volcano-monitoring instrumentation in the United States, 2008. USGS Open-File Report 2009-1165
118. Dzurisin D (2003) A comprehensive approach to monitoring volcano deformation as a window on the eruption cycle. Rev Geophys 41:1–29
119. Benoit JP, McNutt SR (1996) Global volcanic earthquake swarm database and preliminary analysis of volcanic earthquake swarm duration. Annali de Geofisca 39:221–229
120. Power JA, Coombs ML, Freymueller JT (eds) (2010) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston
121. Power JA, Lalla DJ (2010) Seismic observations of Augustine volcano, 1970–2007. In: Power JA, Coombs ML, Freymueller JT (eds) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston, pp 527–552
122. McGee KA, Doukas MP, McGimsey RG, Neal CA, Wessels RL (2010) Emission of SO_2, CO_2, and H_2S from Augustine volcano, 2002–2008. In: Power JA, Coombs ML, Freymueller JT (eds) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston, pp 609–630
123. Neal CA, Murray TL, Power JA, Adleman JN, Whitmore PM, Osiensky JM (2010) Hazard information management, interagency coordination, and impacts of the 2005–2006 eruption of Augustine volcano. In: Power JA, Coombs ML, Freymueller JT (eds) The 2006 eruption of Augustine volcano, Alaska, U.S. Geological Survey Professional Paper 1769. U.S. Geological Survey, Reston, pp 645–667

124. Freymueller JT, Kaufman AM (2010) Changes in the magma system during the 2008 eruption of Okmok volcano, Alaska, based on GPS measurements. J Geophys Res 115:B12415, 14 pp. doi:10.1029/2010JB007716
125. Lu Z, Dzurisin D, Biggs Wicks JC Jr, McNutt S (2010) Ground surface deformation patterns, magma supply, and magma storage at Okmok volcano, Alaska, from InSAR analysis: 1. Intereruption deformation, 1997–2008. J Geophys Res 115: B00B02. doi:10.1029/2009JB006969
126. Larsen J, Neal C, Webley P, Freymueller J, Haney M, McNutt S, Schneider D, Prejean S, Schaefer J, Wessels R (2009) Eruption of Alaska volcano breaks historic pattern. Eos Trans Am Geophys Union 90:173–174
127. Johnson JH, Prejean S, Savage MK, Townend J (2010) Anisotropy, repeating earthquakes, and seismicity associated with the 2008 eruption of Omok volcano, Alaska. J Geophs Res 115. doi:10;1029/2009JB006991
128. Linde AT, Sacks IS (1998) Triggering of volcanic eruptions. Nature 395:888–890
129. Manga M, Brodsky EE (2006) Seismic triggering of eruptions in the far field: volcanoes and geysers. Annu Rev Earth Planet Sci 34:263–291
130. Walter TR, Amelung F (2007) Volcanic eruptions following M > = 9 megathrust earthquakes: implications of the Sumatra-Andaman volcanoes. Geology 35:539–542
131. Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. Phys Today 55:41–47
132. Hill DP, Reasenberg PA, Michael AJ, Arabasz WJ, Beroza GC (1993) Seismicity remotely triggered by the magnitude 7.3 Landers, California earthquake. Science 260:1617–1623
133. Prejean SG, Hill DP (2009) Earthquakes, dynamic triggering of. In: Encyclopedia of complexity and system science, editor in-cheif Meyers RA. Complexity in earthquakes, tsunamis, and volcanoes, and forecast, Lee WHK (ed). Springer, Berlin
134. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1992) Transient stresses at Parkfield, California, produced by the M 7.4 Landers earthquake of June 28, 1992: observations from the UPSAR dense seismograph array. J Geophys Res 100:675–690. doi:10.1029/94JB02477
135. McGee KA, Doukas MP, Kessler R, Gerlach TM (1997) Impacts of volcanic gases on climate, the environment, and people. U.S. Geological Survey Open-File 97-262
136. Robb LJ (2005) Introduction to ore-forming processes. Blackwell Science, Carlton
137. Peterson DW (1996) Mitigation measures and preparedness plans for volcanic emergencies. In: Scarpa R, Tilling R (eds) Monitoring and mitigation of volcano hazards. Springer, Berlin
138. Self S (2006) The effects and consequences of very large explosive volcanic eruptions. Philos Trans R Soc A 364:2073–2097
139. Simkin T, Siebert L, Blong R (2001) Volcano fatalities: lessons from the historical record. Science 291:255
140. Ewert JW, Harpel CJ (2004) In harm's way: population and volcanic risk. Geotimes 49:14–17
141. International Air Travel Association (2010) Volcano crisis cost airlines $1.7 billion in revenue – IATA urges measures to mitigate impact, IATA press release
142. USGS (1997) Volcanic ash – danger to aircraft in the North Pacific. U.S. Geological Survey Fact Sheet 030-97
143. Geothermal Energy Association (2010) Geothermal energy: international market update, 7 pp
144. U.S. Energy Information Administration (2009) Annual Energy Review
145. Ewert JW, Guffanti M, Murray TL (2005) An assessment of volcanic threat and monitoring capabilities in the United States: framework for a National Volcano Early Warning System. USGS Open-File Report 2005-1164
146. Song W-Z, Shirazi B, Huang BR, Xu M, Peterson N, LaHusen R, Pallister J, Dzurisin D, Moran S, Lisowski M, Kedar S, Chien S, Webb F, Kiely A, Doubleday J, Davies A, Pieri D (2010) Optimized autonomous space in-situ sensor web for volcano monitoring. IEEE J Sel Topics Appl Earth Observ Remote Sens 3:541–546
147. Fleming K, Picozzi M, Milkereit C, Kuehnlenz F, Lichtblau B, Fischer J, Zulfikar C, Oezel O, Zschau J, Veit I, Jaeckel KH, Hoenig M, Nachtigall J, Woith H, Redlich JP, Ahrens K, Eveslage I, Heglmeier S, Erdik M, Kafadar N (2009) The self-organizing seismic early warning information network (SOSEWIN). Seismol Res Lett 80:755–771
148. Huang R, Song W-Z, Xu M, Picone N, Shirazi B, LaHusen R (2011) Real-world sensor network for long-term volcano monitoring: design and findings. IEEE Trans Parallel Distrib Syst 99, doi:10.1109/TPDS.2011.170